The Automatic Extraction of Linguistic Information from Text Corpora

by

Oliver Jan Mason

A thesis submitted to
The University of Birmingham
for the degree of
Doctor of Philosophy (PhD)

School of Humanities
The University of Birmingham
July 2006

Abstract

This is a study exploring the feasibility of a fully automated analysis of linguistic data. It identifies a requirement for large-scale investigations, which cannot be done manually by a human researcher. Instead, methods from natural language processing are suggested as a way to analyse large amounts of corpus data without any human intervention.

Human involvement hinders scalability and introduces a bias which prevents studies from being completely replicable. The fundamental assumption underlying this work is that linguistic analysis must be empirical, and that reliance on existing theories or even descriptive categories should be avoided as far as possible.

In this thesis we report the results of a number of case studies investigating various areas of language description, lexis, grammar, and meaning. The aim of these case studies is to see how far we can automate the analysis of different aspects of language, both with data gathering and subsequent processing of the data.

The outcomes of the feasibility studies demonstrate the practicability of such automated analyses.

Acknowledgements

Writing a PhD dissertation is a lonely task, but still relies on the support of a large number of people in the writer's environment. My supervisor, Geoff Barnbrook, has guided me along the (rather longer than expected) way; my wife Joanna has been very supportive and has put up with me spending many an evening in front of the computer; my daughters Freya, Ella, and Rosa have kept me aware that there is life outside the PhD. My father-in-law, Barny Mason, has carefully proof-read the final draft, which greatly improved my written English. My parents, Arnold and Ursula Jakobs, have been supportive and encouraging throughout my whole academic career so far, and I would like to dedicate this thesis to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

For centuries, physicists had believed they knew a vast amount about the world because they knew Newton's laws of gravity and motion. Children at school and students in university learned that they had a grip on the universe because they knew Newton's equations.

So they did, too; after all, they could do all the sums in the textbooks and see that indeed they could work out how a system functioned if they knew the right equations.

However, it turned out that the physicists had never done their sums. Most of what we learned at school is simply not correct. The textbook examples were no more than that: cunning special cases, designed to allow us to ignore friction and other confusions that occur in the real world. The real phenomena are so complicated that they cannot be solved at all, so they were ignored, to allow us to concentrate on a few textbook examples so simple they could be worked out and put in examination papers.

Not until we had computers to do all the laborious calculations for us did we realize that we did not know Newton's laws after all; we had no idea of the confusion, untidiness, disorder, and incalculability they contained.

(Nørretranders, 1998, 357–358)

This first chapter outlines the basic rationale of the thesis. It describes the motivation behind it, and defines the niche into which it fits. It makes the case for automatic analysis, a necessary instrument to drive forward the empirical description of language.

Corpus linguistics has established itself firmly as a methodology for analysing objective data. The in-depth study of individual words and longer multi-word units has contributed greatly to our knowledge of (especially the English) language, with the benefit of more advanced reference works and better teaching materials.

However, so far only a few items have been analysed exhaustively, since this time-consuming task requires a highly-trained specialist. The most comprehensive reference works created using corpora so far have been dictionaries, which have to confine themselves to a description of the most frequent features, because of limited space and time. In corpus linguistics textbooks (e.g. Sinclair (1991), Stubbs (1996) and (2001)) one will find a more detailed analysis, but only of a few sample words or structures. This raises the question of selection: are the chosen examples representative, i.e. will any other word yield a similarly successful analysis of the type described, or have they been selected because they allow easy analysis (which is a perfectly valid reason for an introductory textbook)?

On the other hand, an automated analysis delegates the drudgery of the descriptive work to the computer. Delegation would allow a large-scale analysis of many lexical items across a variety of different corpora, and would shed new light on language in a way not possible at present:

1. lexical items could be compared with each other in full detail, taking more of their properties into account. We could thus compare the collocations of lexical items,

their near-synonyms, the distribution of their inflected forms, their grammar patterns, etc. It would allow us to be more precise when describing the meaning and use of a word in relation to other words.

2. apart from actual choices, potential alternatives can be identified. Looking at language as a series of decisions/choices, it is important to know exhaustively what the possible range of choices is. If we can compare lexical items, we should also be able to define a similarity measure, and with this we can explore alternative choices, and look at whole word fields at once, whereas previously we were restricted to individual items or words seen as alternative choices by intuition.

A number of open questions remain:

what features or properties of words are relevant for the description of language? This will partially depend on the purpose of the description, as different applications require different types of data. Another factor is the possibility of automation; if a feature can be described only with human intervention we will probably not be able to delegate the task to the computer.

how can the relevant features be extracted from corpus data? This includes the question of how much human input is required to get at the information, but also of whether the data contained in corpora will allow us to find the desired information at all.

how can the information be presented in a suitable format? It will have to be user-friendly for human consumption (i.e. dictionaries) or flexible and unambiguous for machine-readable resources. The answer here will probably be a machine-readable format with an extra presentation layer, possibly based on hypertext.

how valid are the results? Relevant issues here are representativity of corpora and volatility, i.e. how quickly changes take place over time, and how stable the results are across different data sets. This is a matter of experimentation with different corpora. It will probably also vary according to the individual feature under investigation.

The last point in particular makes an automated analysis necessary: the procedure will have to be repeated multiple times for different samples of language, systematically varying certain variables. Different genres will have to be compared, as will different regional varieties (e.g. British *vs* American English), or chronologically different samples (e.g. the 1960s Brown and LOB corpora and their 1990s counterparts FLOB and Frown). This is only possible if the procedure is objective and automated and can be performed without too much human intervention.

Stubbs (2003) describes the current situation in corpus linguistics, and identifies the same problem. He states, furthermore, that *corpus linguists have generally been very vague about the methods which they use.* In the work presented in this thesis we aim to make explicit how results are computed, and what possible choices and options were available at different stages of the procedure. Even the use of well-understood terms can be problematic: in an analysis of English morphology, Goldsmith (2001) speaks of a 'corpus' of 50,000 words, but it is unclear what he means by that. He processes individual word forms, so it could be simply a list of 50,000 word types, but if he requires frequency information that could equally well mean a text of 50,000 tokens (and a correspondingly smaller number of types). He does not describe the nature of the corpus in any detail, which leads to difficulties in replicating his work.

Stubbs and Barth (2003) also say that a lot of work is being done on individual lexical items, but that an overarching theory has yet to materialise. The work presented here is an attempt to provide a comprehensive empirical basis for further theoretical work. Obviously this cannot be done fully automatically, but the process of data gathering can be. By 'data gathering' we mean not only the mechanistic collection of material already available, but rather a form of compression: the most detailed map of an area would be on the scale 1:1 (or possibly even larger), but it would be next to useless for navigating. By reducing the scale of the map we lose some detail, but gain an overview (for our purpose) of relevant pieces of information. We are not generally interested in the particular shape of trees (unless used as landmarks), so trees on a road map are either ignored or represented by standardised icons. Buildings are also reduced to stock symbols. In the same way we attempt to create a less detailed representation of the corpus. We abstract away from unnecessary details and try to generalise where possible to summarise what we can find in the corpus. During this process we will throw away a lot of information, and some generalisations will probably go too far, but importantly we will not actually lose anything, since the description we create is separate from the corpus. If possible, links from the description back to the data will allow the retrieval of the original data, similar to the process of navigating through an unknown area with a map.

By producing such a map we will try to discover general underlying patterns, and we will be looking especially for those patterns which also exist outside the domain of language. Such patterns, with the famous example of the distribution described by Zipf's Law (Zipf, 1935), will allow us to link linguistic phenomena to the 'outside world'; they will help us on the way from description to explanation.

We will also need to deal with the distinction between *corpus-based* and *corpus-driven* linguistics (Tognini-Bonelli, 2001): do we work within traditional frameworks of analysis, as most work in corpus linguistics has so far done, or do we abandon the safe waters of established linguistics and venture out into the open sea to discover fundamentally new ways of describing linguistic structures? This question is one that we shall need to return to once the research for this thesis has been discussed.

1.1 Synergies of Scale

> *The language looks rather different when you look at a lot of it at once.*
>
> Sinclair (1991, 100)

Most current research in corpus linguistics is concerned with small-scale phenomena, such as the different uses of formal *vs* informal variants, or usage patterns of particular types of adjectives. Even research focused on broader areas of study rather than individual phenomena is usually restricted to individual examples and requires generalisations based on extrapolations of a few worked examples. It is usually unclear what the status of those sample studies is:

- they could be examples carefully chosen because of some advance knowledge that they would prove a point; they are used in a textbook situation, as other examples do not work all the time or cannot be described as easily.
- they could be 'random' examples chosen to illustrate a point without advance knowledge whether they would work; they worked either by accident or because the methods of analysis apply to all data and always give useful results.

The study of individual phenomena contributes to the overall knowledge of language, and further validates the corpus-based approach, since many if not all of these studies would be quite impossible without a corpus. While they yield useful results for linguistic applications (such as language teaching), their effect on linguistic theory is limited, as long as they are not embedded in an overall model of language.

Without any information on the status of empirical studies it is hard to judge the success of corpus methodology for the description of language. (Sinclair (1991, 53) is a notable exception, as for his analysis of *yield* he explicitly states: *The choice of* yield *is not random, but neither is it chosen because it supports the hypothesis*.) Obviously, corpora have been successfully studied for a long time, in order to answer certain research questions, but a detailed analysis of language usage has not been carried out systematically.

In the structuralist paradigm a sign has meaning only through its relation to other signs, and language therefore has to be studied through synchronic snapshots which provide a complete picture of the system at a given moment in time (de Saussure, 1916). This means that it is nonsensical to look at a sign in isolation and analyse its meaning or function. Through a corpus study of an individual word we can find out some facts about the word, such as what its predominant discourse prosody is, and in what situations it can be used. However, we will not be able to determine its precise meaning and usage without also looking at other words which are used in the same contexts, since the choice between two similar words in a given context (in which both are possible) constitutes a substantial part of the overall meaning. Not saying something that could have been said is almost as meaningful as saying something.

But so far in (corpus) linguistic studies, no truly comprehensive analysis has been performed. The first step, gathering data, has been done for the Cobuild collocations

CDROM (Cobuild, 1995). Here the top collocations for 10,000 headwords have been calculated, but we are not aware of any study of this data base in its totality. This is partly a computational problem. As soon as larger amounts of data are involved it is necessary to automate the analysis. In order to identify similarity classes among the 10,000 headwords one would have to set up a clustering procedure capable of dealing with 10,000 individual items. The next problem is to define the similarity in terms of a metric that can be used to calculate the distance between two headwords in 'collocation space'. These are non-trivial problems which usually require computational resources beyond the means of the average corpus linguist.

However, with increasing computer power, and careful implementation, it should be possible to overcome the problems. Currently the isolated nature of empirical studies is a major obstacle to further understanding of how language works, as looking at a word in isolation means one sees only a small part of the role it has in language. Given that language has to be interpreted as choices against a background of possible options, it is really necessary to have the alternative options as well. The idea is similar to the feature of Cobuild Pattern Grammar (Francis *et al.*, 1996) labelled as 'other words in this pattern/other patterns of this word', which relates each pattern to other words using the pattern, and also lists for each word its possible pattern choices. So, for collocates of a word one should see other words that have similar collocates, since they ought to have similar meaning. Because the analysis is automatic, all the data is available, which it would not be if done manually and on a small number of items. The increase in quantity thus results in an increase in quality, hence the title of the current section. It is a view shared by Halliday (1993, 24), who states that *with the potential for quantitative research opened up by corpus linguistics our understanding of language, and hence of semiotic systems in general, seems likely to undergo a qualitative change*.

Another example of the kind of data that can usefully be gathered for language description is the Longman Grammar (Biber *et al.*, 1999), though here the data is used only for illustrative purposes rather than for the discovery of usage patterns. The corpus used for that project was carefully annotated in order to identify the distribution of linguistic features under investigation.

1.2 Aim of the Thesis

The Path is the Goal (Various)

The aims of this project are to automate the description of language through corpus data as far as possible, and to explore algorithms to do so. There will be several important outcomes, which can be used to determine the success or failure of the work described here:

- We will have created a product, a hypertext dictionary/grammar that describes the language of a corpus (which particular corpus is irrelevant). This language resource will enable the researcher to explore patterns in language on a scale that was not possible before, and where possible similarities will be interlinked and highlighted automatically.
- We will have a software system that can produce such a resource for any corpus (given certain minimum requirements such as corpus size and homogeneity). This system will be an implementation of algorithms that were capable of being automated.
- We will know which procedures could be implemented, and which procedures

required further input data (e.g. of human knowledge) that was not available to the computer. We should be able to determine whether a procedure cannot be implemented at all, or whether it is just the lack of resources that is hindering implementation.

So, while we aim to create a language resource from a corpus, the real aim of this research is not that particular resource, but the development of the procedure to create it. This will involve setting out an inventory of features to be described, how they can be extracted from the data, and how they can best be displayed in a useful way.

The choice of features to be described obviously depends on the feasibility of identifying them automatically, but also on whether they are deemed relevant for our purposes. The intention is not to produce a resource that can be used by a non-linguist, but instead to summarise all available useful information to a researcher (or lexicographer). This should enable the researcher to draw conclusions about usage patterns in the corpus which could hold either generally or for that particular corpus (see section 3.2.2 on representativity for a more detailed discussion).

An important point is that the creation of such a resource should require no human intervention, as it is a computationally intensive task that might take a long time to run on a large corpus. It is clearly not possible in such a situation to stop and wait for human input. That would also introduce questions regarding bias and hinder objectivity. The results of the analytical procedures can be compared only if they have been created under the same conditions, regardless of how any human operator responded to a question thrown up by the system.

We cannot expect the computer to get it right all the time, especially since many procedures involve a stochastic element (such as the parts-of-speech tagger). But as long as the results are consistent and non-random this is not a problem. And agreement on tasks such as parts-of-speech tagging between human annotators is not perfect either; so we have to assume that either human beings are not very good at annotation, or that the systems used for annotation are not suitable for unambiguous application. The latter would provide an argument in favour of abandoning the annotation scheme and following a *corpus-driven* approach.

Without human intervention we will require algorithms from (unsupervised) machine learning and pattern recognition, along with statistical procedures for describing and assessing the outcomes.

The area of research is lexicography in the sense of Quemada (1987) as quoted by Zampolli (1994): Quemada argues for the distinction between 'lexicography', the analysis of lexical units, and 'dictionarics', the creation of dictionaries (what traditionally used to be called 'lexicography'). Indeed one can say that with the advent of computers in lexical analysis the original scope of dictionaries has been extended, and there are now intermediate entities, lexical databases, which can be exploited in ways other than just producing paper dictionaries.

The resulting language resource can obviously be used by lexicographers in the traditional sense to aid the creation of dictionary entries, but it can equally well be used by linguists to study the behaviour of lexical items in text corpora.

1.3 Language Description

What is required for a useful description of language depends very much on the application. Language learners aiming at proficient use of the language will need different information from a computational linguist building a system to translate the language into another one. Thus, the fundamental property of any language description should be *flexibility*. The relevance of different elements cannot be determined out of context, so there should not be any bias inherent in the description. There are, of course, a number of elements so fundamental that they will be useful in most if not all applications.

The central element, which is also used as the primary key for retrieval of the description, is the orthographical form. This can be either a single word or a multi-word unit, comprising either word forms, categories, or a combination of both. Word forms can be specific lexical items or lemmata, and categories can be anything from morphological classes to syntactic or semantic ones. There will certainly be links between each entry and a number of related entries, which will share word forms or categories (or other features, such as meaning or usage in specific genres or text types). For more detailed notes on this see section 3.2.1.2 on page 87.

These links are in fact a further important part of the description: just as a word in isolation does not have (an unambiguous) meaning (e.g. *bank*), a linguistic element can be interpreted only within a context or environment of other elements with which it can be contrasted and combined. In order to understand fully how to use a descriptive adjective, for example, one needs to know which syntagma it can be used in (i.e. what types of nouns it can be used to describe), and what the relevant paradigm is (i.e. what other adjectives can be used to describe the same nouns). Only once this is known can

the meaning be properly appreciated.

This information is not typically accessible through introspection. The only reliable way of obtaining it is through the analysis of corpus data.

The thesis of the work presented here can then be phrased as follows:

> The description of (a sample of) language can be automated to a high degree. Through large-scale comprehensive analysis of linguistic phenomena new insights can be gained which would not be possible with small-scale manual work. Thus automated analysis not only provides a quantitative gain, but also a qualitative one.

1.4 Structure

In the following chapter we will present the research context, the linguistic traditions that we build up on. We will describe the empirical approaches in America and Britain, followed by an outline of the central tool of American descriptivism, the discovery procedure. We will look into the inventory of linguistic units, and assess which of them can be identified with discovery procedures. We will also describe an unsuccessful attempt to discover discourse prosodies automatically, since this will give us some valuable insights into what we can expect from a fully automated analysis.

In chapter 3 we will outline the features chosen for the analysis. We wanted to cover a broad range of features from different areas, namely lexis, grammar, and meaning. We then summarise the methodology that we used for the research presented in this

thesis, including remarks on corpus data and software used for the analysis. Since there is to be no human intervention in the analysis, the software has to take a number of decisions based on statistics. It is important to be explicit about the implementation of the analytical procedures to enable other researchers to replicate studies. For this reason we also briefly discuss matters of preprocessing the corpus data.

The central part of the thesis is a series of case studies presented in the chapters 4, 5, and 6. These deal with various aspects of lexis, grammar, and meaning respectively. In chapter 4 we look into basic statistics that describe the behaviour of a word, such as frequency of occurrence and its spread throughout a text. Collocations are used to give some general indication of both phraseology and aspects of meaning. And finally we will try to describe the phraseology in more detail by looking into ways of extending the single word to arrive at multi-word units. This area is on the borderline between lexis and grammar.

Chapter 5 starts with a brief description of colligation, a somewhat underused variation on collocation which is focused on grammar rather than lexis. We continue with an analysis of commonly occurring grammatical relations, which we call *usage patterns*. These are relations drawn from traditional syntax, such as the relation between subject and verb. We conclude chapter 5 with a more detailed look into grammar patterns, a non-hierarchical approach to the description of a word's grammatical environment.

In chapter 6 we attempt an empirical description of meaning. The first case study compares a (small) number of words at a time and projects their semantic proximity on to a two-dimensional plane based on word co-occurrences. The two remaining approaches pick up the results of earlier procedures, multi-word units and usage patterns, and try to exploit regularities in the data to draw conclusions about the meaning of the

words involved.

Finally, in chapter 7 we summarise and evaluate those case studies, draw conclusions, and outline future work. The case studies indeed show that it is feasible to automate the analysis of text corpora to a large extent, so that the human investigators can concentrate on the more interesting task of making sense of the results.

CHAPTER 2

RESEARCH CONTEXT

In this chapter we will describe the research context in which this thesis is placed. We first discuss the corpus-driven *vs* corpus-based distinction, before briefly summarising issues in empirical linguistics in the past. After that we will review discovery procedures and their role in present-day research; this will cover several areas within linguistics. And finally we will outline the linguistic framework used for the research presented in this thesis.

2.1 Corpus-based and corpus-driven

Tognini-Bonelli (2001, 177) argues *for the establishment of a new discipline within linguistics*, which should be called *Corpus-driven Linguistics*, or CDL. Her argument is based on the distinction between 'corpus-based' work, where existing theories and assumptions are tested on corpus data and 'corpus-driven' analysis, where the starting point is the data, and entities and categories are derived directly from the corpus without being wedded to existing ideas about language.

In this view Geoffrey Sampson would be an exponent of corpus-based linguistics, as he presupposes the applicability of phrase structure grammar for the description of English sentences, but uses corpus data for example to explore facts about the distribution of sentence lengths (Sampson, 2001).

Ideally, all linguistics should be corpus-driven, as the main occupation of corpus-based linguists seems to be adjusting inadequate rules and categories of ill-fitting models to the reality of language. Non-empirical linguists have to rely on their intuitions about language as a source of data, and consequently the range of language phenomena they typically investigate is limited by their imagination and necessarily their idiolect. This makes it impossible to conduct a scientific discussion, as the data used is not objectively verifiable. Any contentious issue can simply be countered by questioning the opponents' examples. Empirical studies, on the other hand, start from the data, and even though it can be described in different ways they have at least got the same starting point. In this kind of linguistics, intuition is used for arriving at a descriptive framework rather than for inventing the object of study.

Obviously, even a corpus-driven linguist in Tognini-Bonelli's terminology initially has to accept some linguistic 'facts' as given, even though these could be rejected later on if they are proven to be either wrong or unnecessary. It would be a step too far to completely reject all existing categories from the beginning, as this would make it impossible to look at the results of new research in comparison with traditional linguistic work. Even if CDL will eventually become the new paradigm in the sense of Kuhn (1963), there still has to be a link to existing work via terminology and/or categories.

In this thesis we will carefully use existing linguistic categories. For example, we will accept the existence of phrases such as noun phrases (NP) or prepositional phrases (PP)

as basic building blocks of sentences, but will not make any further assumptions about the phrase structure of complete sentences. In formal grammar the sentences analysed are usually not nearly as long and complex as authentic ones; and any complexities investigated are mostly due to obscure artefacts of the descriptive formalism used by the analysts.

It is important to keep a certain number of established categories to allow comparability. It would be very difficult to argue for a new approach to language analysis if the units of study bear no resemblance to any existing units such as morphemes or phrases. This might sound more like a political reason than a scientific one, but it is in fact important to allow evaluation of results. Ultimately there is no compelling reason to maintain traditional categories when they have been shown to be inadequate and a comprehensive corpus-driven framework of language analysis has been established. Keeping an open mind does not require one to discard all previous knowledge, flawed though it may be. To start with a blank slate would simply mean ignoring almost everything that has been achieved in language research in the past 50 years.

Another aspect that is worth mentioning at this stage is that a fully automated analysis will avoid some problems with pre-existing assumptions. As long as assumptions have not been explicitly (or implicitly) coded into the software the computer will be free from any bias.

2.2 Linguistics and empiricism

Following the introduction of the corpus-based/corpus-driven distinction, we will now discuss how empirical approaches have been accepted in previous work in linguistics.

Stubbs (1993, 8) lists as the second principle of the British tradition in text analysis that *Language should be studied in actual, attested, authentic instances of use, not as intuitive, invented, isolated sentences.* This principle underlies the neo-Firthian tradition, though not exclusively: Stubbs mentions a number of major American linguists who also base their work on the analysis of empirical data. Both strands will be described in more detail below.

However, the Chomskyan mainstream of theoretical linguistics rejects this principle, as Chomsky himself states: *You want an answer to a non-trivial question, you've got to go beyond looking at data.* (Aarts, 2000, 6). Halliday (1993) on the other hand states that *...data gathering and theorizing are no longer separate activities (I do not believe they ever were)*; this implies that even work on linguistic theory requires looking at empirical data. Sampson (1980, 238) also takes the side of the empiricists, stating that *what makes a theory empirical is a question not of where the theory comes from but of how it is tested. When Chomsky argues that a fully mature scientific discipline ought* in principle *to be treated as answerable to intuitions rather than to observation, fruitful dialogue seems impossible.*

The basic problem is that linguistics can be seen from two different perspectives, as a psychological phenomenon and as a social one. Chomskyans treat language as a mental phenomenon (Green and Morgan, 1996) and thus an area of psychology (and ultimately biology). In this view there is nothing wrong with deriving data by introspection: if language is all in the mind, then any data from outside the mind is obviously less relevant than data from the inside. This point of view allows Green and Morgan to suggest: *Determine (empirically) which sentences are acceptable: ask a native speaker whom you can trust to understand what kind of information you are seeking—often you can act as your own informant*. (1996, 22). They interpret 'empirical' in a very different

way from how members of the language-as-a-social-phenomenon camp would see it.

The other view, that language is a means of communicating meaning between speakers and is thus a social phenomenon, requires empirical data of another sort. It is impossible to make up authentic conversations that reflect realistic use of language. Of course one can hold a soliloquy or script a dialogue, but lacking true intersubjective exchanges they will never be the same as an authentic text sample.

These two views of language cannot really be united, as (according to Green and Morgan 1996) it is the central goal of linguistics to explain language acquisition, whereas empirical linguistics in general tries to explain how meaning is constituted through the use of linguistic elements. Obviously, those are not the only research aims in the two respective groups, but they do reflect the overall direction.

In the following discussion of empirical approaches to linguistics we will take it as given that non-empirical approaches have got little to contribute to the actual study of language, the principal aim of this thesis. We will also not try to reiterate the arguments in favour of an empirical approach; other linguists have already done so convincingly (Sampson 2001, Stubbs 1996, etc).

Schütze (1996) lists four reasons in favour of *Grammaticality Judgments*:

1. to analyse rare constructions which one would not find in a corpus in sufficient number
2. to obtain negative evidence, i.e. about sentences which *do not* form part of the language

3. to separate out *competence* from *performance*

4. to avoid any contextual influence on the language event.

We can easily find counter-arguments which show that grammaticality judgments are not relevant, thereby invalidating those points:

1. describing constructions which occur so rarely that we cannot find examples in a large corpus does not add much to the overall description of language in terms of coverage.

2. given enough imagination one can find situations in which virtually any utterance can make communicative sense; and other sciences can happily proceed without negative evidence.

3. language in its idealised form does not exist; it would make as much sense to study it as it would do to study the movement of objects without taking into account the laws of gravity or aerodynamics.

4. it is pointless to study language without context, as it is an important part of the communicative situation.

Linguistics finds itself in an unusual situation as one of the few sciences where the majority of practitioners invent both the descriptive models/theories and at the same time the data to test them on. But Halliday (1961, 241) affirms that *[d]escription is however not theory*, which rejects the generativists' use of the term 'theory'. Adopting the meaning of 'theory' used in Köhler (1986) one could say that there exist very few (if any) linguistic theories, mainly because linguists rarely use the appropriate scientific rigour

in their studies. Exceptions exist in the form of case studies, such as ones described by Sampson (2001) and Stubbs (2001).

After this initial discussion of the two major views on the epistemological status of linguistics we will now investigate in more detail what the challenges are that one has to face in empirical linguistics.

Linguistics as an empirical science acquires new knowledge through repeated analysis of authentic data, model building and testing, following the ideas outlined by Popper (Okasha, 2002). This will achieve outcomes of sufficient scientific status; the alternative of making up data to test theories and models, a practice often used in theoretical linguistics, simply does not constitute a valid scientific approach to the analysis of language. If linguistics as a discipline wants to be taken seriously, it has to adopt rigorous scientific principles, even if this seems to go against the idea of linguistics as a subject traditionally rooted in the humanities.

One of the basic set of tools in empirical analysis are *discovery procedures*; these procedures are applied mechanically to linguistic data and aim to identify structures in the data without recourse to human intervention. Unlike Sampson (2001), we see empirical linguistics predominantly based in the tradition of distributionalism, rather than simply as a branch of linguistics that uses authentic data to test the models of theoretical linguistics. Discovery procedures will be described in more detail below.

The study of large amounts of language data has to be mainly done by automatic means, as it poses too big a task for human researchers to perform manually (or even semi-automatically). If possible, human intervention should be avoided for reasons of processing speed and objectivity, the latter being important for both comparing different

samples of data and replication of results.

However, the computer has no concept of language, and processing linguistic data thus requires human intervention both at the outset (in selecting and preparing algorithms) and at the end (for interpretation and evaluation). Preparing textual data is non-trivial as it involves a large number of decisions to be made (see e.g. Harris (1985) or Grefenstette and Tapanainen (1994)). Researchers need to take care that decisions made at any point in the processing do not influence the automatic analysis to a degree that it just finds out what the researcher expected it to find.

Ambiguity, which is a common problem in automatic parsing, is not a problem in the reality of language use. Only very rarely is a sentence ambiguous in a given communicative context, and quite often this is a deliberate choice rather than inherent in language. Such sentences are often used in jokes; whereas attempts to construct absolutely unambiguous sentences (e.g. in legal language) end up literally incomprehensible to all but the legally trained. Any description of language should therefore work on the assumption that a sentence has a default interpretation and should avoid creating possible ambiguities. Ambiguities often arise from the use of word classes which generalise too far, whereas a model based on, say, lexical items as basic units would probably provide fewer opportunities for genuine multiple interpretations. However, such a model would be a lot larger, due to the lack of generalisations. The ideal solution ought to be somewhere between the two, a model based on generalisations where appropriate, but with exceptions driven by particular lexical items which behave differently than other members of their respective classes.

Ideally one should use a small number of pre-defined linguistic concepts only. Obviously certain operational definitions prove unavoidable, such as the definition of 'word',

'sentence', 'text', and notions such as 'subject' and 'object'. Word classes pose a difficult problem, as their definitions are generally based on grounds of the semantic, syntactic, and morphological behaviour of a word; often the system of classification originates from a different language such as Latin or Greek. There have been attempts to avoid bias by using numerical labels (e.g. Fries 1952) or to infer word classes from purely distributional behaviour (e.g. Schütze 1993). One can argue about the success and feasibility of these schemes, but eventually we have to judge any classification on its usefulness to the application in question.

The definition of linguistic units by automatic means has been one of the goals of the early distributionalists. Harris (1955) devised a procedure to identify morphemes given a stream of phonemes, though Goldsmith (2001) recently showed that this does not work accurately on larger amounts of actual data. Alternatives (e.g. Creutz and Lagus (2002), Argamon *et al.* (2004)) also do not work fully satisfactorily.

The main reason for the failure of many such procedures to achieve a comprehensive description is that a (natural) language is not consistently designed on independent levels, but has evolved over a long time, coming into contact with other languages, borrowing and discarding elements, and changing into an efficient means for communication that is not at all logical (if such a notion can be attributed to language at all). This is problematic for all purely 'form-based' language processing, as orthographically related forms may or may not be related in any other way (e.g. homographs), and related elements might not share a common form (e.g. irregular forms or derivations).

A further reason for knowledge-poor processing being difficult is that such procedures often ignore higher-level aspects of language, as they usually follow a bottom-up approach only. Such an approach can go a long way, and using more sophisticated al-

gorithms than Harris's overcomes some of its limitations, but ultimately language is a complex interconnected system whose component parts cannot be studied in isolation. Any approach which does not take this into account is bound to have limited results. One example is the development of the success rate of automatic part-of-speech taggers over the past thirty years: after a period of steady increase in accuracy there seems to be an upper limit of about 95% (plus/minus about 3%). Higher accuracy can either not be achieved at all (due to systematic problems with the word class system) or might be limited by the only partial knowledge of structure and meaning that is available for the stretch of text being tagged.

This is one of the major obstacles to linguistic studies: as language is very complex, any description of an isolated component will necessarily be limited in validity, and often has to employ a complex formal apparatus in order to approximate the data (or work with idealised and simplified data). Different syntactic formalisms (for example dependency grammar (Tesnière, 1959) or transformational grammar (Chomsky, 1957)) can be used to describe the structure of many sentences, but like any formal system they are bound to be incomplete according to Gödel's theorem (Hofstadter, 1979). Systems theory, which is suited to the description of self-regulating systems, has only been applied to linguistic description of lexis (Köhler, 1986), which has reasonably well-defined units; it is a lot harder to devise similar systems for syntax or even semantics where the shape of the basic units is less clear.

According to the definition provided by Axelrod and Cohen (2000) we can treat language as a *complex adaptive system*. This would seem appropriate, as there are many forces in language (see Köhler 1986) whose influence can easily be observed for example in morphology. Adopting the complex system paradigm enables us to accommodate those influences, but it also reaffirms the boundary between symbolic and sub-symbolic

processing, which is an additional complication in linguistic analysis. One goal of the analysis is then to untangle the Gordian knot of (sub-symbolic) influences that eventually results in the linguistic form. Failing that we need to at least be aware of the complex nature of language and avoid modelling it with simplistic rule systems.

The power of any formal system to *explain* how language works is also questionable. Pratchett *et al.* (1999) describe experiments in evolutionary computing to design electronic circuits for a particular purpose. The outcome after a number of iterations is a circuit that does what is required, but in a way that no human engineer would ever come up with, as it does not fit with our understanding of how to design circuits. In fact, it is almost impossible to understand how and why the automatically designed circuit works, as its structure is so different from our expectations. But the circuit works, and contains a lot fewer components than a human-designed equivalent circuit would need.

Assuming that language is subject to the same evolutionary forces, we can see that we will have similar difficulties in describing, let alone explaining, how language works. This also means that it is out of the question to use any other than empirical methodology to explore the structure of language. And we also have to abandon any preconceptions about what the structure will look like. This, however, makes it rather hard to evaluate the outcome.

Two major strands of linguistics base their work on empirical data, American distributionalism and the Neo-Firthian British school. We will now look at the two in more detail.

2.2.1 American Distributionalism

Bloomfield (1933) had an important influence on the development of linguistics in America. For the next quarter of a century it was based on empirical principles, the behaviourist paradigm guiding the analysis of linguistic utterances. In this section we will look closer at some of its assumptions (as listed in Wilson (1967, 192)).

- *The description of a language must be based upon a corpus, for instance, the Fries collection of telephone conversations*. This is the same as Stubbs' second principle of British linguistics mentioned above (1993, 2); as such it is the foundation of corpus linguistics, which currently seems to be treated as a branch of linguistics, when in fact it is more a basic methodology. It is not clear how language can be studied in a scientific way without corpus data. Nothing but actual utterances can achieve objectivity of description, and while intuition is permissible when it comes to creating hypotheses, those can only be tested with real data.

- *Any utterance of a native speaker of a language that appears in the corpus must be described and is, therefore, in a sense grammatical*. Considering that actual use as recorded in a corpus is the object of study, there is no reason why the restriction to native speakers should not be dropped. If we view a language as a cluster of idiolects (see section 3.2.2), then it is conceivable that non-native speakers participate in linguistic exchanges like anybody else. Sampson (2001) demonstrates that there is no clear distinction between grammatical and ungrammatical sentences. The question of whether this blurs the overall description needs to be answered with reference to the actual purpose of the analysis: linguists working on a dictionary for learners might choose to ignore corpus data which could cause confusion.

A comprehensive description, however, will have to account for everything that is significant in frequency.

- *There is (or it is possible to develop) a mechanical procedure for revealing the grammar of a language [...]*. This is the underlying assumption of the thesis set out at the beginning of this text. It is also important for any large-scale analysis of linguistic corpora, as computer processing is necessary to cope with large amounts of data. The question remains how far such procedures can go at present, but in principle there should not be an upper limit. Obviously it depends on the definition of 'grammar': if it includes all aspects of language (which it did not in the American structuralists' view, where it was limited to phonology, morphology, and syntax), it might be problematic when dealing with meaning. Sampson (2001) states that meaning is outside the scope of empirical work; other people (e.g. Rieger (1989) and Teubert (1999)) disagree with that.

- *The importance of a structure may be judged by the frequency with which it occurs*. This is also an important principle, as it allows us not only to make statements about the significance of certain observations, but also to introduce thresholds to ignore rare events. As language phenomena frequently consist of a large number of rare events (Baayen, 2001), this is important for computational purposes. This principle also tallies with another principle of British linguistics, namely that *much language use is routine* (Stubbs, 1993, 2). Routine implies repetition, and thus increased frequency. This applies not only to word combinations, but also to non-lexical phenomena. As language relies on a limited set of common elements shared between speakers (vocabulary, structure, etc), these elements need to be used over and over again, and are therefore important. Rare events often result in communication problems (e.g. with unknown words); unless these can be re-

solved through recurrence to a known (frequent) environment the utterance has missed its purpose.

- *Language is binary. Any structure is divisible into two immediate constituents [...]*. There does not seem to be any convincing argument that provides a basis for this claim. On the contrary, when using phrase structure grammar to describe the structure of sentences (which is a prime example of 'immediate constituents' analysis) the restriction to binary branching leads to complicated and inelegant structures full of 'dummy' non-terminal nodes. Dropping this constraint leads to more direct descriptions which can do with a smaller number of non-terminal symbols. Occam's razor should be applied in this case; should a compelling reason be discovered at any later stage it is always possible to transform a non-binary analysis into a binary one anyway.

While some important principles of Bloomfieldian structuralism are still valid, others seem to be less appropriate now: the restriction to phonology, morphology, and syntax, for example, excludes lexis and meaning; and more recent work has led (in the British tradition) to the principle that *form and meaning are inseparable* (Stubbs, 1993, 2). Another result of much work in lexis is that lexis and syntax cannot be studied independently, whereas structuralists worked with the *additional assumption that one must not mix levels; for example, the syntax must not be called in to help describe the phonology* (Wilson, 1967, 192). Sampson (1980, 223) states that *this controversy is now quite rightly a dead issue*, illustrating this with an example where the concept 'word' is necessary in a phonemic description, thus breaching level boundaries.

However, despite some unnecessary over-restrictive principles the structuralists turned linguistics into something more of a science. Unfortunately this trend was later reversed

by the mentalist approach advocated by e.g. by Chomsky (1957).

With Harris American structuralism moves on into its second phase, distributionalism (Helbig, 1983). Now the aim of linguistic study is to identify units (i.e. segment the data), group units with shared environments into classes, and describe the distributional properties of these classes. 'Discovery procedures' are used to identify the units; they are purely mechanical and explicitly exclude meaning. However, meaning is permitted as a useful 'shortcut', to reach faster a conclusion which could be reached more slowly through mechanical procedures. As meaning and distributional environment are equivalent (two units occurring in different environments will have different meanings) this would yield the same results anyway. The introduction of meaning into the analysis, however, weakens the empirical position of distributionalism, especially since short cuts were often unavoidable due to limited amounts of available data. Helbig (1983, 83) also gives the counter-example of colour terms, which generally occur in similar environments, despite having different meanings. However, this seems to be based more on speculation than on actual research; colour terms will surely share some environments, but will typically be used in some individual contexts as well. Certain colours (e.g. *blue*) have meanings that go beyond the purely descriptive 'colour'. And not every item can have every colour, so some restrictions exist here as well.

According to Helbig (1983) there were few practical results of distributional analysis, and Harris moved on to transformational analysis. However, the lack of useful outcomes can easily be explained by insufficient amounts of data and lack of sophisticated methods. In the recent past distributionalism has been revived, especially within computational linguistics, and the general availability of large corpora together with advances in information theory, machine learning, statistics, and other related areas have shown that distributionalism is a feasible approach to the study of language. This means

that criticism of early distributionalist approaches needs to be re-evaluated and that one cannot dismiss them on the basis of arguments that were only valid half a century ago.

2.2.2 British Contextualism

Having discussed American structuralism as an essentially empirical branch of linguistic history we will now have a look at developments in Britain, where a different school of structuralism, *contextualism*, was introduced by J. R. Firth (1890-1960). After first looking at a few key issues we will contrast contextualism with Saussurian structuralism, as they differ in some important aspects.

Stubbs (1993) gives an account of the principles underlying much of British linguistics in a Firthian tradition. Even though the American structuralists approached linguistics from an anthropological perspective, they nevertheless viewed language as a formal system, unlike Firthian linguistics which concentrates more on the social function of language, treating it as embedded in society and culture (Helbig, 1983). The orientation on empirical analysis of use (i.e. *parole*) is shared, but in general the focus seems to be more on 'high-level' (i.e. social) phenomena, rather than a mechanical analysis up to the level of syntax.

Possibly motivated by the large vocabulary of English, and the resulting number of near-synonyms, in the mid-1960s 'lexis' was introduced as a new field of study. Firth had been working on it earlier, but arguably it was established as a field in its own right by Halliday (1966) and Sinclair (1966). Previous approaches had treated words simply as fillers for slots within syntactic structures, defined through word class categories; the individual choice of words would be determined by semantics (with certain grammat-

ical restrictions) and thus was outside the scope of analysis. The focus on the actual words in a text on the other hand led to the discovery of basic mechanisms (e.g. 'collocation', see Sinclair *et al.* 2004) for the patterning of language. It also made obvious the interdependency of lexis and syntax, that certain words pre-select certain syntactic constructions and *vice versa*. It was therefore pointless to study syntax and lexis in isolation, which in consequence generated new approaches to the description of grammar (e.g. Hunston and Francis (2000)).

When he laid down the foundations of modern linguistics, de Saussure (1916) introduced a number of dualisms. Unfortunately, many of these dualisms do not add anything to the study of the subject, but instead compartmentalise it into subparts with no real existence. In the remainder of this section we will discuss two of these dualisms, 'system *vs* instance' and 'synchronic *vs* diachronic', and we will argue that neither of them is relevant in the Neo-Firthian approach to language. The point of this discussion is to emphasise the differences between contextualism as practised in the UK and *bona fide* structuralism.

2.2.2.1 System *vs* Instance

The first dichotomy is that of *langue* and *parole*. Being a structuralist Saussure postulated a language system, which governs the way language is used, just as the rules of chess describe how to play the game. These rules are independent of the actual games played, so the latter are basically applications or projections of the rules. In language *langue* describes the structure of language as such, independent of any actual usage. *Parole* on the other hand is exactly that usage which is based on *langue*, i.e. it follows its rules, but exists independently.

Parole would thus be the object of study of empirical linguistics, while *langue* would be the underlying model to be explored through the study of *parole*, though in modern contextualism the distinction has mostly been abandoned (see below).

A literal use of 'system' is made by a model of language introduced by Köhler, which treats language as a dynamic system in the context of synergetics and systems theory. The model is based on functional relationships between system components and can therefore be seen as equivalent to *langue*, and it can be tested on *parole* data. This model identifies relevant variables within the system of lexis, and states functional relationships between them. These relationships are formulated as (differential) equations, so that the model can actually be tested on corpus data. So far the model has been successfully tested on German (Köhler, 1986) and English (Gieseking, 1993) data.

Chomsky (1965) later introduced a slightly different dichotomy, *competence* and *performance*. While *performance* is practically identical to *parole*, *competence* is in the mind of the idealised 'native speaker', rather than an external abstract system as *langue*. Unlike Saussure, who had not connected any evaluative judgments with his pair, Chomsky uses his dichotomy to define the object of study of linguistics: *competence* is what is important, as the actual utterances of a speaker cannot be used to derive any information about their linguistic knowledge. This is due to many influences from the 'real' world which are completely unrelated to language. He categorically denies the validity of any frequency counts, as those would merely reflect sociological facts, e.g. that New York has more inhabitants than Dayton, Ohio; for this reason he rules out corpus analysis as a valid source for linguistic data, as it will be incomplete and skewed (Chomsky, 1957).

Leaving aside the problem of having a mental construct as the object of scientific study, the 'hard sciences' have long had to deal with imperfections in their experiments,

and the laws of gravity have been described despite the influence of air resistance on the actual speed of a falling object. Chomsky's objections to the study of actual usage thus do not need to be disproven in detail, since they are obviously flawed from an empirical perspective: there is no objectivity possible if the 'data' is invented by the researchers themselves for no good reason.

Going back to Saussure, Stubbs (1993) has described in detail why even the distinction between *langue* and *parole* is unnecessary. As we interpret an individual utterance against the background of our linguistic knowledge (based on previously encountered utterances), they are effectively the same thing. Stubbs mentions an analogy (originally introduced by Halliday) of weather and climate, or micro and macro. For the study of language neither the Saussurian nor the Chomskyan dichotomy carries any significance at all.

One basic problem with the idea of a uniform model/system is that no language is completely homogeneous, as every speaker will have their own (internal) 'version' of it, and a language is essentially the sum of all its idiolects. In fact, Pilch (1976) asserts that the object of study in linguistics is not language, but texts; languages, being a theoretical concept, cannot be observed, only texts can. When we analyse a text we assume that the text has been composed in a certain language, but Pilch points out that this is not always true, as multiple languages can occur in any given text. Leaving aside lexical items, which can easily be transferred into other languages, either by morphological adaptation (e.g. German *checken* and *gecheckt* from English *to check*) or directly (*a priori*, *zeitgeist*, or *en route*), this also applies to syntax, where L2 speakers may introduce structures from their own L1 into texts composed in their L2. These transfers are not easily recognisable, as they do not concern the form of items but rather their (sequential) arrangement, and while it is fairly easy to keep words from different languages

separate, that is not nearly the case with something less accessible by conscious thought like grammatical structures.

One could even postulate that there are no languages as such, but instead that human beings speak a number of idiolects which are similar enough to each other to allow understanding. Instead of the idealised native speaker of Chomskyan linguistics we thus have to deal with a multitude of individual variations. The aim of linguistic description would then be to describe the common core, or overlap between these idiolects. However, if we consider that languages/idiolects form a continuum we might find that sometimes there is no overlap at all, that two people both see themselves as speaker of language X without being able to communicate. This could be predominantly the case with more distant regional variations, e.g. Bavarian and Frisian in Germany (though they might be equally incomprehensible to any non-local speaker of 'standard' German and thus count as languages in their own right). It does seem strange to abstract from these variations to an idealised version of a language (which typically will be the 'educated' variant spoken by linguists, who are concerned only with their own intuitions). Therefore we cannot rely on non-empirical data for a realistic description of language, and we have to realise that language is not a simple entity, but has multiple facets instead.

If we adopt the view that languages as such do not really exist, but instead are amalgamations of idiolects, then being able to speak several languages simply boils down to having multiple idiolects; 'German' would be an idiolect chosen for talking to another speaker of German, while 'informal English' is the idiolect chosen for conversations with friends, and 'highly formal English' is reserved for appearances in court or academic lectures. All these idiolects necessarily co-exist in the speaker's mind, and are thus susceptible to interferences, sometimes leading to 'mixed' dialogues (with elements from

different idiolects) Such dialogues are perfectly possible, which suggests that all languages/idiolects are processed in the same way rather than compartmentalised. Their borders are thus far from clear-cut, and make idealisations difficult.

This, however, leaves us with the problem that our object of study (i.e. language) simply disappears, as we cannot objectively define it. What we effectively do is to study 'fossilised' language, samples that have been recorded in a particular communicative situation. In order to make sense of the data we need to be aware of the situation, as we can derive information about the kind of language used from it. This need for contextual meta-information raises the issue of balance and representativeness of a corpus, which we will discuss in more detail below (see section 3.2.2).

Rieger (1989, 21) introduces the term 'pragmatically homogeneous' for text corpora if they contain utterances produced in similar contexts/situations. This provides a useful short-cut, as any situational variables would remain constant, so that we can use a simplified model of language as an average over multiple idiolects, and in practice we need not worry about the finer philosophical problems, provided our corpus is somehow controlled for variation.

To summarise: for an empirical study of language it is important not to make any artificial distinction between authentic data ('instance') and an underlying theoretical construct ('system') which is supposedly the object of study. The study of language has to focus on actual utterances, and any patterns that can be identified are initially only valid for the data analysed, and can only be taken as generally applicable when they have been found in other samples as well (see section 3.2.2). This reflects the dependence on the communicative context of an utterance, and the notion of homogeneity will need to be applied to the data being investigated.

As to the matter of whether languages exist, we will in this thesis use the term 'language' to refer to the union of idiolects that are represented in a homogeneous sample of utterances. The notion of frequency of occurrence can then be used to give us an idea whether a certain phenomenon is widely used or rare. Should the corpus consist of texts by a single author, then we would in fact analyse that person's idiolect. What we can usefully say about language in general from such a limited sample remains to be seen; but common sense suggests that the overlap between idiolects must be reasonably large in order to enable communication between speakers. Only the repeated analysis of many idiolects can eventually tell us what the degree of overlap would be.

2.2.2.2 Synchronic *vs* Diachronic

Another of Saussure's dichotomies is that of *synchronic* versus *diachronic*. Whereas the modern linguist concentrates on the *synchronic* study of language, i.e. the 'snapshot' taken of language at the present time, not taking into account any change over time, the historical linguist is concerned with *diachronic* analysis. In practice this distinction is rather meaningless, as it is not possible to sample language without a time dimension. Any data is collected at a certain time and place, and as the speed of language change is not yet sufficiently explored, we cannot decide which time differential would be small enough to ignore. A language changes gradually, and changes propagate through time (Renouf, 2000) and space, though not necessarily at the same speed.

Because of modern mass communication a language will be more homogeneous now than it would have been in the past, but there are still geographical factors which are probably just as important as the chronological ones. It is therefore far better to face up to the reality of the nature of language than to hang on to pointless idealisations.

That means we have to take into account that no language is a homogeneous entity that remains stable wherever it is in use, and that (regional) variations change at different paces.

One of the consequences this has for empirical work is that the results are valid for a limited time only, and updated versions of the same corpus data will need to be examined when available. This makes desirable the existence of monitor corpora (Sinclair, 1996a) which are constantly updated during their lifetime. However, while actual studies are affected by such diachronic changes, the methods of empirical analysis discussed in the following chapters are not. For that reason the distinction between synchronic and diachronic linguistics has no influence on the claims set out in this thesis.

2.3 Discovery procedures

The very term 'discovery procedure' seems to have been invented by Chomsky as a stick to beat structuralists with. Monaghan (1979, 48)

'Discovery procedures' meaning mechanical methods of linguistic analysis without recourse to meaning have been seldom used other than with languages where there is no native speaker doing the analysing. In the American literature, they had a temporary vogue, but they were never a feature of the British tradition. Monaghan (1979, 40)

We can define a *discovery procedure* as a 'mechanical' way of identifying linguistic features. The use of the word 'mechanical' implies that a discovery procedure does not rely on any human input during its application; once set up it will work without any

further intervention by the researcher. The early distributionalists developed a number of such procedures in the past and other researchers have updated them more recently. In this section we will comment on their usefulness and whether they warrant inclusion in the system described in this thesis.

Historically discovery procedures are mainly linked to Zellig Harris, who describes how to move from phonemes to morphemes (1955) and from morphemes to utterances (1946), though, according to Nevin (1992), Harris himself did not believe that discovery procedures alone were sufficient for language analysis. Fries (1952) can be counted as a discovery procedure for word classes, and more recently there have been attempts to improve on Harris' work which will be described below in more detail.

In his *Syntactic Structures* Chomsky (1957) originally sets out a programme for deciding which of two grammars would be more appropriate. This is based on a 'discovery procedure', which takes as its input a corpus, and produces as its output a grammar. This procedure is then *a practical and mechanical method for [...] constructing the grammar* (1957, 50–51). A 'decision procedure' then takes a corpus and a grammar as input and decides whether the grammar is the best grammar for it, and finally an 'evaluation procedure' takes two grammars and a corpus, and decides which grammar is the better one.

The first stage in the empirical analysis of language is the *taxonomic* one, where individual items are identified with the aim of subsequently describing their relations to each other. One fundamental problem is the physical reality of language: naturally, it exists only in the form of sound waves, which are hard to describe without the aid of sophisticated (digital) signal processing technology. In an ideal world linguists would still regard spoken language as the 'purest' manifestation of language, whereas in prac-

tice most researchers work with the more pragmatic choice of written language, which is far more tangible. Here the most basic units become immediately obvious, partly due to orthographic conventions:

- the letter (grapheme), which is seen as analogous to the transient and elusive phoneme,
- the word (a combination of letters separated by spaces or certain kinds of punctuation),
- the sentence (terminated by corresponding punctuation marks), and
- the text (a complete 'utterance').

All of these units are a lot harder to find in an acoustic stream, where co-articulation and other factors reveal phonemes as a convenient but non-existing abstraction, and words run into each other without any obvious pause between them.

Written language is a lot easier to process, even though the conventions of spelling are not unambiguous either. The various roles of the full stop (as a termination of an abbreviation or a sentence) and the apostrophe (as a single quote, an elision marker, or a possessive marker in English) cause real problems in the automatic processing of language, but it is still a lot easier than dealing with spoken language.

Assuming the existence of the abovementioned units, even though they are simply derived from the written form, there are clearly units in between them, which are not as easily identified. These would be morphemes, which are components of words, and phrases, which combine to form sentences. Their existence is postulated since it is ob-

vious to the observer that there are regularities on sub-word and sub-sentence levels which ought to be describable more specifically.

Morphemes are defined as the minimal units in language which carry meaning. This definition has multiple flaws, in that it is not objective (as meaning cannot objectively be defined), and imprecise. Traditional morphological analysis does provide some plausible cases, but also some difficult ones.

In many cases meaning can only be assigned to morphemes on the basis of etymology, which is basically educated guesswork (and is often wrong in the case of folk etymologies). Frequently morphemes on their own have no sensible meaning whatsoever, and a synchronic description could not take diachronic developments into account anyway. In the case of words derived from other languages, knowing the meaning of their components in the original language does not really contribute much to knowing the overall meaning of the words; once they have become English words their meaning has developed independently from the original elements. This also becomes apparent when looking at the full etymology of a word like *receive*, where Latin *capere* turns into *cipere*, is combined with *re-*, becomes Old French *recoivre* and finally (ca. 1300) is adopted into English as *receive* (URL, 2005) (A slightly different etymology is given in Collins (1991)).

If splitting words into morphemes is to be attempted it clearly ought to be an objective procedure. Ideally the (empirical) linguist should work without any intuitive preconceptions/abstractions/ideas about how language works, as that would undoubtedly influence the outcome of any study. Instead one would use a discovery procedure to identify relevant units or divisions of units that follow some general principles rooted in universals which are usually imported from areas outside linguistics. Examples for

this are Zipf's Law (Zipf, 1935), Menzerath's Law (Altmann, 1980), or—as a method—minimal description length (MDL, Barron *et al.* 1998).

However, even though such procedures typically perform reasonably well, given that they are completely void of any linguistic knowledge, there are some fundamental problems:

Collecting linguistic data already abstracts from sound to phonemes in a way that could not easily be 'discovered' using an objective procedure. This is perhaps the one situation where one can accept Chomsky's objection to studying *performance*, without accepting the dichotomy itself: speech is so variable, and influenced by so many factors that it would require an enormous amount of data to draw any statistically significant conclusions. No two sounds uttered by any speaker are the same, due to physiological constraints, and even supposedly identical sounds are modified by their phonetic environment to a degree that they do not share many acoustic properties. Judging from research in speech recognition it generally seems an utterly unfeasible task to use discovery procedures to segment a continuous stream of acoustic signals into phonemes.

One could object that the human child does exactly this when acquiring language in its first years. However, a human baby typically has correlated non-linguistic input (mainly visual) which makes the discovery of sounds a lot easier. Also, it is questionable whether a child actually acquires phonemes; it seems much more likely that some larger units of sound (e.g. syllables or even whole (short) words) are acquired first, and that the phoneme is basically an abstraction reached via the influence of an alphabetic script which has no actual existence in language. This seems to be corroborated by observing young children as they learn to read and write: they do not usually break up words into phonemes, but other arbitrary units.

So, in the case of spoken language it does not appear to be possible (yet?) to apply discovery procedures to 're-discover' phonemes, and extra-linguistic input would also be required to provide additional stimuli/input. With written language the main problem is the comparative arbitrariness of the writing system(s) used. Especially in the case of English there is no clear correlation between phonemes and graphemes. Instead the writing system is an amalgamation of representations of multiple features of language; and where in speech a phoneme can be realised by a number of allophones depending on context, in writing the combination of higher-level units affects the spelling of words, as morphemes can similarly be realised through allomorphs.

In fact, the traditional definition of a morpheme as the 'minimal unit of language that carries meaning' is hard to defend when one accepts that meaning is not only conveyed through lexis, but instead arises from a combination of different units in a particular environment of higher level units. If meaning is not restricted to morphemes, then why should all morphemes carry meaning?

There is also a problem with the combination of units: often the units themselves change, in speech through co-articulation, and in writing through doubled consonants or elided letters. These phenomena make it a non-trivial undertaking to identify the units, since suddenly there is no longer a one-to-one correspondence between types and their related tokens. Even if it is possible to identify 'allo-units' in a text, it is not possible to automatically assign them to a common abstract unit, as they usually have complementary distributions.

There is a further fundamental problem with the use of discovery procedures, namely that it is impossible to evaluate the results in the light of current linguistic knowledge. Language has to be initially viewed as a 'black box', as we cannot get access to its in-

ternal workings, and linguists in the past have attempted to create a possible internal structure of such a box in the form of rules and formal systems. The success of these attempts is debatable: rules rarely function without exceptions, and we are a long way away from comprehensive coverage of utterances and their structure. This point is similar to the evolutionary development of language as mentioned above.

Now, trying to identify linguistic structure by automated procedures one is faced with the difficulty of dealing with a holistic entity. Even though language has in the past been divided into areas such as phonology, morphology, syntax, etc, this division is purely artificial and research (e.g. in speech recognition) increasingly shows that certain problems cannot be resolved without recourse to higher levels of description. Ultimately we have to deal with language as a whole. Therefore, the linguistic model of the black box is just a model, and using discovery procedures we will undoubtedly get a different model, especially if they start off with little or no initial linguistic input. How can we possibly decide which of the two models is correct, or even better than the other?

Corpus-based work in phraseology seems to indicate that utterances are made up of chunks, prefabricated blocks which are used over and over, without being 'grammatical' units in themselves. Householder (1982, 288) introduces the analogy of improvisation in (jazz) music, where musicians use such prefabs which are *stretches of some length*. He furthermore quotes related work (Sudnow, 1978) in sociology which suggests that it is more than a mere analogy, but points to *an essentially identical neurological mechanism* (Householder, 1982, 289). Units of meaning as computed by Danielsson (2001) fall into this category of prefabs, as do the collocational frameworks of Renouf and Sinclair (1991). It is highly unlikely that discovery procedures will produce items that correspond to current/traditional linguistic units, i.e. that comply with rules or expectations of grammaticality.

Evaluation is thus difficult if not impossible, as we have no benchmark to evaluate the identified items against. We would have to resort to general measurements, such as coverage and size of the description, and we would have to look at them without any preconceptions. This will be discussed in more detail below in section 4.3 on multi-word units.

In the following sections we will briefly summarise and evaluate current research using discovery procedures (though they are not necessarily called that). This survey will be structured according to the traditional subject areas within language description.

2.3.1 Discovering Morphemes

From a Saussurian structuralist point of view, morphology is a mess. Morphemes are the smallest units which are not overly influenced by physiological constraints (as is the case with phonemes) and which are very frequent. As a result they are heavily influenced by dynamic processes. These processes enforce a number of optimality conditions (see Köhler 1986 for an application of these principles on lexis) which lead to a constantly changing inventory. Worse still, the pace of change is not even, words have different 'ages' and therefore will have been subjected to changes for differing amounts of time. Taking a synchronic 'snapshot' we cannot tell which words are how old, and that makes analysis very hard.

In addition, words are constantly imported and exported between different languages, and therefore morphemes from other languages interfere with 'native' morphemes in that they bring with them their own rules (e.g. with plural formation). This makes it difficult to rely solely on the criterion of meaningfulness for the definition of

morphemes.

Furthermore, morphology being at the boundary of lexis and phonology leads to a number of other influencing factors. Co-articulation changes individual phonemes, but also adjusts morphemes when they are combined, leading to allomorphs that complicate matters. This applies both to pronunciation and spelling.

Even though morphemes are traditionally defined as the smallest units that carry meaning, there is no consistency, and any link between a structure and its meaning is met by a number of counter examples. A multitude of examples can be retrieved from compounds, where (in German) *Hustensaft* is medicine to combat a cough, while *Orangensaft* is juice made from oranges. Similarly in English, *shoulder bag* and *leather bag* have the same morphological structure, but very different correspondences between the meaning of their constituent morphemes and overall meaning. This is a general problem of linguistic structures, in that they usually have multiple interpretations which are disambiguated either by context or by the actual elements occurring in the structure.

Despite all those difficulties, morphemes have been the subject of one of the best-known examples of discovery procedures: Harris (1955) demonstrated how morphemes can be identified from a sequence of phonemes by looking at the transition probabilities after each phoneme position, i.e. the number of choices available to continue the sequence. At each local maximum (in the sequence of probability values) he postulated a morpheme boundary, which did actually work on a sample sentence. Harris originally worked on a phoneme-string, but in principle the algorithm should also work on graphemes.

One problem that would occur, though, is the increased possibility of misclassifica-

tion, as the mapping from phoneme to grapheme loses information due to the smaller number of graphemes as compared to phonemes. For example, *unity* and *uninformed* have distinct prefixes phonemically, but identical ones graphemically. This is a case where the loss of information blurs a relevant distinction between the morpheme *uni-* and the string *un-*.

Goldsmith (2001) lists further examples where Harris's algorithm goes wrong and suggests that the problem lies in its inability to distinguish between two types of variation, *freedom due to phonological combination* and *freedom due to a boundary between two morphemes*. In other words, we have here a mixture of phonological and morphological influences, and it is difficult to identify which of the two combined influences (or possibly even both) causes the degree of variation. In a similar situation in digital signal processing one could use a Fourier transformation to decompose a complex signal into the sine-waves (of different frequencies) that, when added together, result in the signal. It would be useful if there were an equivalent solution to disentangle the various forces (analogous to frequencies) which simultaneously act on the utterances we are investigating.

Another weakness of Harris's algorithm that Goldsmith identifies is that it operates on local criteria only, and does not take into account the language as a whole. By finding a description that is oriented towards global criteria individual mistakes can be avoided. We could probably say that Harris's algorithm easily identifies 'local optima', whereas we are looking for a 'global optimum'. And for that we need to look at all words and morphemes at the same time.

Goldsmith uses an approach based on minimum description length (MDL) to automatically identify morphemes. His morphological model is rather simple, though, as he

only uses ‘stem+affix’ to describe the morphological structure of a word. Both Creutz and Lagus (2002) and Argamon *et al.* (2004) combine MDL with a more complex, recursive model, where a word can be split into (binary) constituents, which can in turn be split again. Fine-tuning of the cost functions used for evaluating inventory size and the number of morpheme boundaries in a word can be used to calculate an optimum morpheme inventory from a given corpus. If the parameters are not well chosen one ends up either with 26 ‘morphemes’ (the letters ‘a’ to ‘z’) or at the other extreme with mainly mono-morphemic words.

The algorithms described are also vulnerable to processing order: Argamon *et al.* (2004) explain the meaningless morph *-ter-* which is derived from *inter-*; previously *in-* had been identified as a morph. An implementation of the algorithm described by Creutz and Lagus (2002) suffered from similar problems. In general the output of such procedures is mostly reasonable, but contains a number of problem cases, which are often caused by orthographical conventions (e.g. when combining certain morphs) or by the unrelated repetition of character sequences. Productive affixes with no spelling irregularities are identified with good success, but a fully automatic discovery of morphemes seems to be impossible.

Perhaps this is the limit of (empirical) morphological analysis: identifying productive combinations where both elements occur either independently or bound to other elements, while ‘frozen’ combinations are not analysed further. Words would then be atomic in the case of *analysis* and *dialysis*, as neither of the components *ana/dia* and *lysis* exists independently, whereas *swim+s* remains analysable, as *swim* exists as a free morpheme. The word *internationalisation* would be *inter+nation+al+isation*, reflecting the kernel *nation* and the recurring productive affixes. The final suffix could be *is(e)+ation*, though this full analysis could probably not be identified automatically,

unless some information on spelling adjustments are programmed into the algorithm, which is then no longer a proper (mechanical) discovery procedure.

In summary we can say that there are some promising attempts, but no fully satisfactory solution. It seems to be the case that the limits of automatic discovery have been reached, and even fine-tuning of parameters cannot improve the result significantly. This could mean either that the current algorithms are not good enough, or that morphology is too unsystematic to be analysed in such a way. The latter option would imply that a discovery approach would not be feasible in morphology.

2.3.2 Discovering Word Classes

The traditional word class system that most of Western linguistics has been based on is derived from the work of scholars describing the Latin/Greek language (Covington, 1984). As a result, it does not fit any other language that is even moderately different in structure. This became clear during an exercise in the PAROLE project when a common tagset for the languages involved was going to be devised. Interestingly enough, English, the most frequently analysed language in modern linguistics, did not fit at all into the pattern of feature values set out for the other European languages.

Word classes are usually defined using a number of ill-fitting principles. None of those principles works properly on its own, and there is typically a large number of exceptions. This raises the question whether words can usefully be assigned to word classes at all, given the number of ambiguities and the variability of words, especially in English, where most nouns can also be used as verbs.

Principles used to classify words include:

Meaning: words referring to objects are nouns, those referring to activities are verbs, and properties are given by adjectives. (One undergraduate student of English thus classified *resignation* as a verb, since it describes an activity).

Morphology: word forms which combine with the same inflectional or derivational morphemes share the same word class. This, however, only applies to nouns, verbs and adjectives in English.

Function: adjectives can modify nouns, adverbs can modify adjectives. If the dependency relations are known we can derive the class of a word by the words it modifies or is modified by.

Distribution: words can be classed according to the syntactical environments they occur in. For example, a word fitting into the environment *the ??? car* cannot be a verb, only an adjective or another pre-modifier.

Fries (1952) describes in detail why the current system of eight to ten parts of speech is flawed; his main argument is that it does not have a single basis for the classification, but mixes arguments from form and function in a completely unsystematic way. He instead proposes a system based on substitutability, where 'class 1' words are those that for example fit into the frame '(The) ___ is/was good'. The system comprises 4 classes (roughly equivalent to the traditional classes noun, verb, adjective, and adverb) and 15 groups of function words. All word class assignments are based on possible positions in a test frame, which would in principle allow for automatic discovery, given the test frames.

However, it also requires a 'seed' list of words from the classes, as Fries starts off with a very specific frame (see previous paragraph). Only a very limited number of class 1

words does actually occur in that frame, especially if one is looking at a corpus where sentences tend to be much more complex. At later stages, as more classes and groups have been introduced, he moves on to sentences which contain few lexical items and consist almost entirely of class/group labels. If these frames can be matched in a corpus (which is more likely) then unknown words could be classified according to the class that would be required in the position they occur in.

The other problem is that Fries' system requires human input to judge whether the resulting sentence is structurally the same. There are some example sentences in linguistics which show that surface similarity does not correlate with structural similarity: *Warren is eager to please* has a very different structure from *Warren is easy to please* (Winograd, 1983, 138).

Fries also discusses the formal characteristics of word classes, and here he gives further details which can be used to identify them. The emphasis here is on 'identify', not 'define': it is a property of a class 1 word to follow a group A word (such as *the* or *my*), or to have a morphological variant with an added '-s' (*boy* and *boys*). But these properties were not relevant during the derivation of the word class system, instead they have been reached *a posteriori*.

A more recent approach is Schütze (1993), who uses context information to cluster words into classes with similar distribution. His classes seem to be more fine-grained and show elements of both grammar and meaning that contributed to the grouping. The problem here is to set the thresholds for class inclusion, and to find meaningful labels for the classes. Furthermore, each word type is only ever a member of a single class, which is problematic for homographs.

Steiner (2004) also addresses the issue of automatic word class identification and hits on the rather interesting point that there is a mixture of syntactic and semantic features which seem to influence many automatic procedures. We will look into this aspect of word classes later (section 6.1.2) when we examine the meaning of words.

To summarise: there have been several attempts to derive word classes, which seem promising overall. However, fully automated procedures like the one by Schütze tend to have a large number of classes, which makes them hard to use for grammatical descriptions. Fries' approach uses fewer classes, but is not as easy to implement without human involvement.

Another question is what word classes are going to be used for in the first place. In principle they should be an abstraction for the formulation of grammatical rules, in which case a smaller number of classes would be more useful. But a smaller number of classes might cover up finer distributional differences. It is impossible to determine *a priori* what the 'right' number of word classes should be.

2.3.3 Discovering Phrases

The aim of discovering phrases in text is to investigate what larger contexts a word is used in. It is not to analyse complete sentences in terms of their syntactic structure, but instead to focus on the individual lexical item and its syntactic environment. Clearly it is desirable to analyse larger chunks of language as well, but here the focus remains on the word.

2.3.3.1 Some notes on Syntax

Before the subject of phrases is covered, a short section on the status of syntax is necessary. Most of the mainstream work in syntax seems to be looking for the Holy Grail of an all-encompassing formalism, which describes all the sentences of a language which are grammatical, and marks those that are not. This has been set out as the task of linguistics by Chomsky (1957). While pursuing this goal, most syntacticians got sidetracked into solving intricate problems of little importance to the actual user of language. Computational approaches (e.g. Garside *et al.* 1987, Black *et al.* 1993) achieve reasonably broad coverage with statistical methods, but have to compromise with a reduced depth of analysis.

However, Gödel's Theorem (Hofstadter, 1979) states that any axiomatic system in mathematics is either incomplete or contradictory, and if this applies to a carefully constructed artificial system (mathematics) one can safely assume that it is also valid for a biological, evolving system (human language). Therefore it seems highly unlikely that the 'perfect' grammar will ever be created; and Sampson (2001) has shown that there is no clear-cut boundary between grammatical and ungrammatical sentences. Similarly, Atwell (1988) argues that a comprehensive phrase structure grammar of (unrestricted) English would be too large to be of practical use. A survey of other rule-based parsing systems shows that most of them do not nearly have enough coverage despite pushing computational capabilities (from 20 years ago) to the limit. Atwell's grammar, derived from a sample of a treebank comprising about 2000 sentences from the LOB corpus, has 8,500 rules, too many for the Prolog system he uses. He also observes that there are many obvious gaps in the rules, where structures have simply not occurred in the sample. Therefore it seems futile to waste any effort on constructing a comprehensive

grammar, but what else should be the goal of syntactic description? Undoubtedly, sentences do have a structure, and for the vast majority of utterances it is not too difficult to describe.

The answer lies in a shallow and lexicalised approach. It is comparatively easy to identify constituents (or chains of dependent words) with fairly minimal effort based on traditional word class assignment. In those cases where the word class's behaviour is different from the behaviour of the word token in question, a lexical approach (e.g. Gross 1997) can be used to capture idiosyncrasies.

So, this section is based on the assumption that a complete-coverage grammar for full sentences is unachievable, but that instead sentence fragments can be safely identified. The interrelations of those sentence fragments are then a different problem which will be discussed in section 5.3.

2.3.3.2 Discovering Phrase Structure

In analogy to the morphological discovery procedures described by Harris (1955), Tretiakoff (1973) presents an algorithm to derive IC phrase markers from transition probabilities of word classes. He calculates the ratio of the conditional probability of tag_j given the preceding tag tag_i and the independent probability of tag_j using the formula

$$C_{ij} = \log_2 \frac{P(tag_j|tag_i)}{P(tag_j)} \tag{1}$$

This he calls 'degree of correlation', but it is in fact quite similar to the mutual information score (Church and Hanks, 1989). The tokens with the highest correlation value

are then combined until all tokens have been joined up in a phrase marker. He gives the following example (1973, 218):

```
WORD CLASS    PROBABILITY   TOKEN

 44                         SHE       ***
             2.564                       *****
 01                         LOVED     ***   *
             1.232                           **
 45                         A         ***   *
             2.379                       *** *
 05                         GOOD      *** * *
             1.860                         ***
 04                         LAUGH     *****

Word classes:
01 - transitive verb indicative
04 - common noun
05 - qualificative adjective
44 - personal pronoun
45 - indefinite article
```

Tretiakoff concedes that he only worked with a small sample of 3500 words (200 sentences of a novel by Somerset Maugham, of which 72 were correctly analysed).

In order to see whether this is a suitable method for deriving syntactic structure without any preconceived human input that could be biased by existing theories (apart from the obvious bias that the structure can be represented by a phrase marker, that the tree is binary branching, and that each token can be assigned a certain word class), a parser was implemented following Tretiakoff's description. As a data base for the calculation of the probabilities the transition matrix of a stochastic part-of-speech tagger (QTag, see Tufis and Mason (1998)) was used. This matrix has been derived from one million words of tagged data. Using this matrix somewhat shortcuts the procedure as outlined by Tretiakoff, as the probabilities can be easily calculated from it.

Running the parser on the example sentences given in Tretiakoff's paper gives the identical structure, though unsurprisingly with different probability values. A further

difference is the use of a slightly different tagset, but this would not invalidate the method as such.

Looking at the example sentence *What is wrong with his morals*, the two implementations produce the following results:

1. (((What is) wrong) (with (his morals))) [Tretiakoff]
2. (((What is) wrong) ((with his) morals)) [Mason]

Tretiakoff's analysis is closer to the traditional view of grammar, as it identifies the sentence final noun group. In the replicated analysis this is split, as the possessive pronoun has been grouped with the preposition. One counter-example obviously does not invalidate the method as such, which clearly warrants further work (if a binary tree structure is indeed the aimed for target of the syntactic analysis, see section 2.2.1). But it makes interpretation rather difficult: again the problem is one of input. Using a discovery procedure to derive syntactic structure does not lead to labels such as 'noun group', which are only available when using preconceived models of analysis. With morphological analysis this is different, as we are not primarily interested in whether a morph is a stem or a suffix, but rather in the inventory of morphemes. However, with syntactic structure it is the label as well as the grouping that is required to make most use of the result in subsequent processing steps.

Overall the results with a Tretiakoff-style parser look very promising, especially when considering that they are only derived from (empirical) bigram transition probabilities. One of the strong points is that the parser has complete coverage without the need to hand-craft a grammar: any sentence will be parsed, even 'ungrammatical' ones, and as

the structure is based on distributional principles it will rarely be completely 'wrong'.

Evaluating the success or failure of an automatic parser is an almost impossible task. Sampson (1995) mentions a workshop held in 1991 at a computational linguistics conference, where researchers were given the task of annotating a set of sentences in labelled bracketting. The overlap in analyses was minimal, even though the target was simply the identification of constituents in what Sampson thinks were *not ... unusual or specially problematic English constructions* (1995, 5), and the language used was English, probably the one language that has attracted most research effort in analysis so far. He traces the disagreement back to the lack of explicit standards or norms, and continues to present a proposed standard in the form of his SUSANNE scheme of analysis.

Again we have the dilemma of comparability. While it might be possible to set up evaluation tasks for given applications (for example SensEval (Kilgarriff 1998, Edmonds 2002) for word sense disambiguation and Parseval (Black *et al.*, 1991) for parsing), this is not the case for (exploratory) empirical basic research. We do not know what the outcome will be, so we cannot in advance tell what it is going to look like. The procedure by Tretiakoff described above might well be one that delivers adequate analyses in terms of immediate constituent identification, and we cannot judge its results using traditional ideas about phrase structure. All we can form an opinion about is the general principles involved in the algorithm, which seem plausible enough.

2.3.4 Discovering Meaning

When talking about 'meaning' it is typically the meaning of words that comes to mind first. It is word meanings that are codified in dictionaries, and grammar is thought to

simply put these meanings into relations, adding functional roles such as subject and object. In linguistics this is often refined by using morphemes rather than words as the principal units that possess meaning, and 'meaning' is extended to include linguistic features such as 'plural' or 'past tense' as part of meaning. However, the problems with the meaning of individual morphemes as described above remain.

In their introduction to semantics, Chierchia and McConnell-Ginet (1990) only get to word meaning in chapter 8 (out of 9). They refer to Frege (1884)'s *context principle*, which states that words *only* have meaning within a proposition, a view also expressed later by Wittgenstein (1921) in his 'Tractatus Logico-Philosophicus': *3.3 Only the proposition has sense; only in the context of a proposition has a name meaning*.

Along similar lines to Frege, Gross (1982, 297) argues that *the smallest unit of meaning is the simple sentence*, regarding word meanings as implausible. Traditionally, function words are considered devoid of any meaning, and Gross wants to extend that notion to content words as well. This would suggest that words are not the correct units to use when trying to locate meaning in language. In Gross' view words need to be contextualised, and he uses bare sentences providing the minimally required surrounding words to do this. The meaning of a transitive verb thus depends on what is chosen as its object; without the object it has no meaning.

Despite this opinion, many linguists (and lexicographers) still view lexical items as primary carriers of meaning, though increasingly the influence of the environment (both grammatically and lexically) is acknowledged as important to constitute meaning. A word may have a 'basic' meaning, but this can be enhanced or restricted through co-occurrence with other words and grammatical constructions. A word can also form part of a larger unit, such as idioms or larger sequences, where it can completely lose its

original meaning. Obvious examples include *to kick the bucket*, where neither *kick* nor *bucket* have their usual meaning, but also *not visible to the naked eye*, where the word *naked* is not used in its literal meaning, but rather means 'unaided', warranting its own entry in Sinclair (2001).

Sampson (2001, 206) states that *word meanings are not among the phenomena which can be covered by empirical, predictive scientific theories*, as (word) meanings are too subjective, fuzzy and ill-defined. The idealised abstractions of predicate calculus and logic cannot be applied to a reality that is essentially messy, contains no absolutes, and is always changing. *Semantics cannot be scientific* (Sampson, 1980, 237), due to the *unregimented, unpredictable working of the conscious human mind* (p. 236). This view is basically correct, and it is indeed hard to think of a way in which two unconnected systems, that of language and that of external reality, can be kept synchronised (which would be a precondition for objective analysis). It does not matter here whether one operates in the context of Cartesian dualism or Popperian pluralism: there are always at least two worlds, and entities from one cannot directly be linked to the other; mainly because the physical world is singular, whereas the world of mental objects and events is unique to an individual, and thus not common to all of humankind. However, what is perfectly possible is to map the (internal) relationships between lexical items based on an operationalisation of 'meaning'.

As Stubbs (2001, 35) states: *The vocabulary of a language is not an unstructured list of words*. So, while we cannot empirically describe the meaning of a particular word as long as it requires some referential link leading outside language, we should be able, by investigating text corpora, to find out about the internal structure of the vocabulary, the relationships between individual words *within language*. This, to a human observer, ought to provide us with a meaningful description of the word's usage, which is after

all one possible definition of meaning. Even if links between the two systems 'language' and 'outside world' cannot be established, the internal structure should be isomorphic, i.e. if 'outside' items referred to are similar in some form, then the referring 'inside' terms should also be similar. This is supported by research into aphasia, where Huber (1981, 423) observes that *knowledge of the surrounding world is typically better preserved than is the linguistic capacity to give a name or a semantically well-formed description.* So there seem to be two distinct representations of the world in the brain, one linguistic and one 'encyclopedic', which would remain unaffected by dysfunction of the linguistic one.

Obviously, we enter here the territory of individual interpretations of reality, and there will be multiple perceptions of reality in different texts, and consequently multiple structurings of the vocabulary, each isomorphic to its corresponding perception. Unless we are dealing with texts originating from a single author (and created within a short time span) we will get multiple interpretations overlaying each other in the data. For the computational analysis this translates as a lot of statistical 'noise', and we will only be able to approximate a mean of the different individual meanings. This is conceptually similar to an experiment that Elman (1995) describes.

It clarifies matters if a distinction is introduced between the *referential* meaning of a word and its linguistic one. Danielsson (2001) gives the example of the (Swedish) word *atonala* which occurs five times in her corpus. From the five concordance lines it is possible to infer that the word is used to describe a style of music, but it remains unclear what style it is exactly. The closest we get to realising the meaning of the word is through contrast with other words that we know can be used in such situations, or in this case through a morphological clue: it looks as if *atonala* is the opposite of *tonala*, which might be a known word. So, we can only move within the domain of language in

our search for meaning, and the result is the *linguistic meaning*.

The *referential meaning* of a word, on the other hand, is what it refers to in the outside world; it is not a part of language, but rather a part of what is sometimes called *Weltwissen* ('world knowledge'). The *linguistic* meaning of a word is grounded in language; it describes e.g. what other words it can be used with in a linguistic relation. So, without knowing exactly what 'atonal music' is, we can identify it as a meaningful combination, whereas an 'atonal plate' does not make immediate sense (though it might do in a specific sublanguage context where 'plate' has a technical meaning). Pilch (1976, 91) also states that *[t]he study of referents is outside the province of linguistics*.

Philosophers have tried to wrestle with this though, notably Russell and Wittgenstein. However, a detailed account of their studies is outside the scope of this thesis.

One problem with referential meaning is that it is far too variable to be usefully constrained, mainly due to the creativity and flexibility of human language use. Hastings (1994) gives the example of CAMILLE, a system for automatic incremental acquisition of lexical meanings, using a set of constraints for possible syntactic objects of verbs. He gives the example of *torched*, which has the constraints `actor = terrorist` and `object = building`. These are used in connection with a syntactic analysis of the corresponding sentence to identify the (logical) object as an instance of 'building' and the (logical) subject as a 'terrorist'. However, a brief inspection of a few concordance lines of *torched* shows that the object can include vehicles and human beings, and the subject can be anything from mobs to soldiers to criminal gangs to opponents in a game of cricket, even a volcano in one case. A system working on assumptions given by the above constraints would not be able to analyse many of these sentences correctly, due to the high degree of variability and creativity in language, part of which is captured by metaphorical usage.

In defence of CAMILLE it needs to be said that it is designed to work in a subdomain of terrorism texts, but providing it with this information brings no actual knowledge gain, and even in specific subdomains language can be used creatively.

Pedersen and Chen (1995) describe a system which works along similar lines; like Hasting's CAMILLE their system requires a concept hierarchy to form generalisations. The example given in their paper sounds very plausible, but operates on a set of very basic sentences, which would not typically be found in authentic language data.

Manning and Schütze (2000, 294) call the acquisition of meaning the *holy grail of lexical acquisition*. Many applications in natural language processing could greatly benefit from including meaning, but there are also difficulties of representing meaning in a way that a computer system can make use of. In many cases the acquisition task is then re-cast as identifying semantic similarity instead, in order to deal with unknown words or to allow generalisations. This practice is in line with the above argument of two distinct/unconnected worlds, since identifying links between language and reality is too complex, whereas intra-linguistic relationships can be determined automatically.

By looking at corpus data it ought to be possible to gather information about the linguistic meaning, whereas referential meaning is usually outside the scope of corpus analysis. Nevertheless, using techniques from the area of terminology extraction (e.g. as described by Pearson 1998 the identification of referential meaning might sometimes be possible, e.g. when definitions can be identified in the text (i.e. patterns of 'a NOUN is a NOUN', which exemplifies the lexical relation of hyperonymy, and can be part of a definition). Further examples for the discovery of lexical relations using a pre-defined set of patterns are Girju *et al.* (2003) for meronymy and Hearst (1992) for hyponymy. For the automatic derivation of referential meaning from dictionary definitions see Barnbrook

(2002). However, this then raises the issue of knowledge representation, i.e. how this information is to be represented within a formal system.

The accepted view in corpus linguistics is that the meaning of a word is in its use; this is primarily based on the work of Firth, but also mirrors the contextualist view in the philosophy of language (Frege, Wittgenstein). Each word (or other linguistic element) is interpreted against the background of all available options/choices, so by analysing certain contexts we can observe what items occur in those contexts, and by keeping the context constant we can see what possible alternatives exist. We can then postulate a similarity between the possible options, as they can be used in the same (linguistic) context.

The regularity between form/usage and meaning is employed by Ruge (1997) to find (near) synonyms to use as possible search term extensions in information retrieval systems. Starting from a set of basic dependency (head-modifier) relations, words are seen as similar if they share similar modifiers. If an object has a property 'size' it is likely to be described by adjectives expressing 'size', and if it has a property 'colour' it will be modified by colour terms. Words which share a lot of properties (of their linguistic meaning) will typically refer to similar objects in the real world. The results of their work seems to confirm the assumption that there is a structural similarity (isomorphy) between the linguistic and the referential meanings, i.e. that language somehow mirrors the real world. The relationship between these two meaning types is not objective, but rather exists as a mapping in the minds of the individual speakers of a language. There is obviously a lot of overlap between the mappings, otherwise communication would be difficult; other problems include synchronisation, i.e. language adapts to changes in the real world through a shift in the referential meaning (which might distort the relationships between the linguistic meanings), or vice versa (when language is deliberately

used to influence attitudes and perceptions of reality).

In the philosophical discussion on theories of meaning this model of a structured vocabulary would be subsumed in the holistic approach, where each element's meaning depends on the global configuration of the network of all elements, and any change in the meaning of one element leads to changes in the meaning of all other elements. While some philosophers use psychological and common sense arguments to ridicule this view, it makes perfect sense if one does not require identity of meanings for the purpose of communication. Two people can talk about a topic as long as their internal definitions of the concepts used are sufficiently similar, and it is not relevant that minute adjustments to meanings are continuously taking place.

2.3.5 Discovering Discourse Prosodies

Extending the discussion on whether discovery procedures are suitable to identify meaning, we will now describe a further application where discovery procedures are difficult to apply. This reinforces the argument from the previous section that there are two distinct 'worlds', and that links between the two cannot be established by automatic procedures.

2.3.5.1 Definition and Examples

Louw (1993) presents a phenomenon termed *semantic prosody*, which captures the effect of a lexical item which one would normally read 'between the lines': the prosody of an item indicates an aspect of the speaker's judgment or opinion, and thus the author can use it for dramatic effect in poetry, for ironical remarks, or for the (unintentional)

marking of insincere statements. If the actual environment of an item clashes with its expected prosody it causes such effects.

An example (from a Birmingham undergraduate student essay on linguistics) is the statement that certain linguistic phenomena *are endemic in this text*. The word *endemic* has a clear negative prosody: in the BBC corpus it co-occurs with words such as *fear*, *violence*, *war*, *disease* and *corruption*. This prosody is part of the meaning of *endemic*, and any collocational analysis will show this. Hence the word is not applicable to linguistic entities, unless they are judged as negative, and even then it is rather strong.

Stubbs (2001, 65) proposes alternative terms of *pragmatic prosody* and *discourse prosody*, and decides to use the latter, as it emphasises the way the prosody contributes to creating a coherent discourse.

Discourse prosody focusses on a 'higher' level phenomenon, the shared properties of the collocates of a word: if a word has predominantly 'negative' collocates, then it has a *negative prosody*, and the interpretation of any 'positive' words in its environment needs to take this into account. Stubbs (2001, 45) gives the example of *CAUSE*, which has a negative prosody, as one can see from the following random set of concordances from the FLOB and FROWN corpora (different data from that which Stubbs used):

```
ostered by the Iraqi authorities may cause further misery for the Kurdish peopl
ortem examination failed to find the cause of death. Professor Malcolm Lader, o
that the advertisement was likely to cause considerable offence if it appeared.
sible that the reporting season will cause the market to falter. The two key fa
its first overall deficit since 1967 cause a rising tide of members to quit. "C
 in budgeted sponsorship is the root cause of its present deficit. Lord Sainsbu
house for a year." This approach may cause difficulties, both with conductors a
uld be made "gradually, so as not to cause unnecessary hardship." * Parliament
ing crew. Dover coastguards said the cause of the tragedy was still unknown. Bu
```

20 minutes to control the blaze. The cause of the blaze is still being investig
in Guildford Place early today. The cause of the blaze is being investigated.
y and fume-filled and buses using it cause congestion in Percy Street. They wan
of the Theatre Royal. New cash rules cause row THOUSANDS of council-house tenan
she concluded: "This proposal would cause unacceptable harm." President opens
it then becomes an all too frequent cause of regret at having spent the money
eas one in 200 (report, May 18) is a cause of considerable concern, and should
ployment. But the unemployed did not cause inflation. Ministers now have to dem
be legislated for that are likely to cause the most noise in the run-up to the
ns had been killed demanded that the cause of death should be recorded as "buri
form of dividends. There are other causes for concern about GEC's recent recor
howing the pain education sometimes causes children. Sue Prescott, one of the p
appeals are perhaps one of the main causes of a drop in income among some chari
only makes it appear necessary, but causes slight shrinkage, thus averting the
s that it is not incompetence which causes accidents but showing off. The real
nt improving other hostels but what causes concern is that the policy of the Yo
e of burying the past. Wrong number causes a few red faces George Parker's WMN
. It is the sin of the priest which causes him to turn their dignity into disho
the degree of camera shake detected causes image frames of your subject to be s
through the city where it obviously causes a great deal of attention and provid
trategy that tackles the underlying causes for the rise of Le Pen. Whether the
our, I shall examine the nature and causes of the rising concern about judicial
al society and in understanding the causes which had led to it, whereas Durkhei
precision when discussing symptoms, causes and conditions. People who are not f
her. In trying to outline different causes and effects, we have focused on what
aken in churches except for special causes. Weekly collections were introduced

One can obviously see some exceptions (mainly with the noun usage, due to the phrase *good cause(s)*), but overwhelmingly the examples show negative events. Stubbs then looks at the examples of the more specific *CAUSE + amusement* he finds; these can not be taken as counter-examples, but rather describe amusement at someone else's expense. The positive aspects of amusement get overridden by the negative sense of *Schadenfreude*. Here the discourse prosody allows the speaker to express a more detailed assessment of the situation through the clash of the negative and positive expectations of the two words involved. Similarly, Louw (1993) provides an example where a careless

choice of words betrays a speaker's real intended meaning: he uses a negative prosody when talking about a supposedly positive event. This kind of analysis suffers of course from the impossibility of reading a speaker's mind, but a (competent) informant would notice the oddity of such an unusual combination of words.

2.3.5.2 Problems of Automatic Identification

We can link discourse prosodies with semantic concepts, such as 'positive' and 'negative'. In the absence of the computer's capability to judge events, a human informant will need to provide the procedure with some initial data. This could, for example, take the form of a list of negative words and another list of positive words. The computer could then check the environment of a node word for occurrences of words from either of these lists; and from a simple frequency count of the proportions of positive and negative words in this environment it could infer the type of prosody. The computer could then tentatively add any unknown words encountered into the category with the higher frequency, provided there exists a significant difference in frequency counts. We might well find that some cases are inconclusive when no category predominates.

For operational purposes we can then re-interpret discourse prosodies as proportions of membership in a number of categories: if out of 100 concordance lines 27 contain 'positive' words, 59 contain 'negative' words, and the remaining 4 lines contain no known words, then we can talk of a 60% tendency towards a negative discourse prosody. We can thus apply *fuzzy set theory* (Zadeh, 1965) when dealing with discourse prosodies.

One important caveat applies here: all judgments will remain subjective. While almost every speaker will probably classify some words, such as *catastrophe* or *accident*, as

negative, other words derive their evaluation from cultural, ideological, and individual preferences. For example, Darwinists will judge *the decline of the creationist tradition in Western culture* as positive, whereas orthodox followers of Christianity might not exactly agree with this assessment. Similarly, *profits* sounds good to advocates of a free market economy, but rather bad to socialists. As a consequence, no objective criterion exists to decide on the 'correct' discourse prosody of a given word.

This dilemma results from the ambiguous nature of language, 'caught' between society and the individual. While we view language as a social phenomenon, necessarily intersubjective to allow communication between individuals, it still originates from those individuals without any 'global' system of rules.

As stated in the previous section, we cannot identify the referential meaning automatically from text corpora. Discourse prosodies however refer to the outside world; nothing inherently positive or negative exists in language. One possibility would be to start from a 'seed' list of positive and negative words, and then exploit the patterns in which these words are used to explore the prosodies of other lexical items. We could achieve this through an iterative stochastic procedure.

While this method of evaluating discourse prosodies seems to yield reasonable results in some cases, a closer analysis brings to light some serious problems with a simple word-based approach. First, we can not decide whether a single word has a positive or negative prosody in isolation. Leaving aside issues of *homonymy* and *polysemy* the context of use can effectively invert the 'stand-alone' lexical meaning. For example, for the node word *decline*:

- ...and a 10.0 percent decline in 1994 but was well below its 35 percent rise in

1989...

The word-based discourse prosody algorithm identified this as a *positive* example, as the list of positive words contains *well*. Similarly, *high culture* is 'positive', but in the context

- fears about the decline of high culture

fears is followed by a negative phrase, as *decline* gives a negative 'spin' to it. So, while *decline* in itself would count as predominantly 'negative' in isolation, it seems by and large to have a 'neutral' or even 'positive' discourse prosody. In cases where the associated element has a negative prosody, as in *a sharp decline in IRA terror activities* the overall prosody turns into a positive one; so the phrase *fears about a decline in IRA terror activities* [I] would exhibit what Louw (1993) calls 'a clash' showing irony or insincerity, or simply a rather unusual view of British politics.

So, a purely word-based approach will not achieve enough precision. Instead, a possible improvement could include a phrase-based approach, which involves evaluating a complete phrase at a time. This tallies with Leech (1974), who postulates the existence of intermediate 'units of meaning', as no lexical items exist in English for certain concepts, such as Leech's example of *young monkey* as opposed to *young cat/dog/chicken/...*

Another possible model one could investigate would involve some kind of 'semantic predicates'; we can then model the word *decline* as:

```
PROSODY(DECLINE(X)):
    not PROSODY(X)
```

Here we define the prosody of *decline* as the logical negation of the argument's prosody. Used with a negative prosody (*decline of terrorist activities* [I]) this results in an overall positive prosody, whereas a positive prosody (*decline of disposable income* [I]) ends up as an overall negative prosody for the whole phrase.

2.3.5.3 Conclusion

The previous section has shown that there is a limit to what we can find out about language automatically. As soon as we leave the domain of language itself and try to connect linguistic items or structures with external values or meaning we encounter severe problems. The reasons for these problems are easy to identify:

1. meaning is based on an internal representation which cannot be linked in an objective way with reality.
2. language is not confined to a single person, and thus value judgments attached to individual elements are not generally valid.
3. vagueness inherent in language makes it difficult to automatically transfer properties from one word to other words, even if they are deemed to be similar (for example through a shared environment).

It is perfectly possible to investigate discourse prosodies, and several researchers (e.g. Louw, Stubbs, Sinclair) have already done so; however, this is one aspect of language that is beyond automatic detection.

2.3.6 Bootstrapping

If we want to be faithful to an empirical discovery of linguistic structures we have to start with a clean slate, i.e. without any preconceptions about the structure of language which might bias the final outcome before we have even started the procedure. However, starting at zero is very difficult and time-consuming, even though it ought to be possible (if one discounts the existence of a Chomskyan LAD). If children can learn to use language, so should the computer (eventually). But we need to make a distinction between applying a structure (which does not require explicit knowledge of its categories) and knowledge of that structure. One can perfectly well use a language successfully without knowing anything *about* the language, and knowledge of all the meta-information does not, on the other hand, mean that one can actually use the language.

To describe a language we do require meta-information, i.e. a set of categories and rules governing their interaction. There are two basic ways of getting such categories: deducing them from the data (slow and error-prone) or presupposing them in advance (fast and error-prone). Because categories that we are going to identify automatically are bound to be fuzzy, data-driven discovery will need careful tuning of thresholds etc, while prescribed categories will generally be too 'crisp' to fit the observations properly. In consequence, we will need to find an adequate compromise between the two.

2.3.6.1 Ideal and Real World

As shown through the evaluations in the previous sections, there is still a long way to go before discovery procedures can be used to guide the analysis of language without

any human intervention. This is mainly due to the lack of progress in linguistics since the late 1950s (Paprotté, 1992), when truly empirical work was pushed to the sidelines by the mainstream which was focused more on intuition.

In an ideal world the present study would be based entirely on such discovery procedures, but in order to achieve the aims of this work it is necessary to take a number of 'short cuts', to bridge the gap between the ideals of theory and the problems of application. That is not to say that it will not at some point be possible to rely purely on discovery procedures, but at the present stage they need to be 'helped along' in order to produce useful results. In the process decisions will have to be made, and the outcomes of the procedures will need to be guessed where steps are missing. Undoubtedly an element of error will be introduced, as it will not be possible to anticipate the results without actually running the discovery procedures on real data. However, we will attempt to keep the risk to a minimum by not putting too much emphasis on categories that have simply been assumed.

It is important to stress that this does not amount to 'cheating', as the procedures will still not involve any human input apart from *a priori* information and categorisation, which one can replicate using similar procedures. Once this has taken place, the procedures described here can be re-run with a modified set of basic categories or units, and they will then produce valid though different output.

We need to be careful, though, to avoid any bias being introduced. Fries (1952, 8) states that *[a]s a general principle I would insist that, in linguistic study and analysis, any use of meaning is unscientific whenever the fact of our knowing the meaning leads us to stop short of finding the precise formal signals that operate to convey that meaning*. He instead advocates the use of purely structural and formal properties, though guided by

meaning as a means to test whether two structures are identical or different in meaning. While this test for semantic equivalence is hard to implement by computer, Fries' general principle can easily put into practice using computer programs.

In the following chapter we will discuss in more detail the methodology applied in this thesis, describe the computational and statistical procedures employed, and the data used for testing purposes.

2.4 Summary

In this chapter we have argued that the only feasible approach to a description of language is empirical. Language needs to be investigated in context, and in examples of real use, as many of the potential ambiguities can be excluded in authentic examples. Stubbs (2001) gives the example of *surgery*, which has four main meanings, but for any instance of it one or two words besides it are sufficient to determine which of the meanings is the appropriate one.

We have seen that early approaches to language study were essentially empirical: the American distributionalists were analysing authentic data, and so have linguists within the British contextualist tradition. The distributionalists were discredited through the Chomskyan mentalist re-orientation, which rejected authentic data as spoilt in favour of introspection. The British tradition continued, but often did not actually use authentic data (Stubbs (1993) mentions examples in Halliday's work which are obviously invented).

The introduction of computers and the increased availability of large scale corpora

has sparked new interest in distributionalist ideas, and has also given the contextualist approach a boost, which lead to notable successes, such as the Cobuild range of dictionaries and grammars. This coincided with the realisation in the language engineering field that the results of theoretical linguistics are not useful for applications, which often do better using statistical approaches.

To conclude, there is a new optimism in the field that empirical methods are the way forward, though many empirical linguists are not aware what is possible to achieve with automated methods. It is the purpose of this thesis to investigate how feasible it is to apply fully automated methods to the analysis of corpora, and to advance the possibilities of describing features of language on a large scale without any human intervention.

Chapter 3

Methodology

Following the outline of the research context, we will now investigate three main areas of language in order to work out the feasibility of a fully automated analysis of corpora. Those areas are

1. Lexis
2. Grammar
3. Meaning

These areas are taken from linguistic tradition; but they have never been more than rough delimitations of subject areas. Language is a holistic phenomenon, and its study therefore often has to cross artificially imposed boundaries. Therefore these areas are not to be taken as absolute, and some studies will indeed be hard to assign to a single one of them; However, most studies will usually be predominantly be located in one specific area.

The main reason for choosing those areas is that they have been in the centre of attention in recent work in corpus linguistics, so they should be more advanced than other areas. Furthermore, the current project only works with text corpora, rather than acoustic data. For that reason phonetics/phonology is left out. Areas beyond semantics (e.g. pragmatics) focus on individual contexts/situations, and are thus not as easily accessible through generalised corpora. These areas would belong to *text analysis*, rather than *corpus analysis*, and they require a fundamentally different approach.

3.1 The Problem of Choice

The first problem when selecting a number of case studies is which to choose. There is a large number of potential cases to analyse, but for reasons of space (and time) only a small number can actually be studied. In this section we will briefly list which areas will be covered in subsequent chapters. After that we will discuss the methods used for implementing those case studies.

Hoey (2003) lists five questions that one should ask about a word:

1. What does the word mean?
2. What words does it associate with?
3. What meanings does it associate with?
4. What grammatical functions does it associate with?
5. What positions in the text does the word favour?

We will attempt to cover all these questions in this thesis. In the following sections we will now describe its overall outline.

3.1.1 Lexis

The chapter on lexis starts with the analysis of general distributional properties of word forms. The properties chosen are those that a) contribute to the description of the word form's behaviour, and b) can be retrieved automatically. They include frequency of occurrence, spread of occurrences, inflectional variations, and use of tense, aspect and voice.

We cannot answer Hoey's final question with general corpora, as we do not always have access to the individual texts of such a corpus, or the positions of a word within the texts. But in principle it would be trivial to answer if that information was available. As an approximation, we could attempt to describe the nature of the overall distribution of a word form: is it evenly spread through the data, or does it occur in clusters? This would give an indication of whether the word is a specific or general one.

Sinclair (1991, 44-51) describes an analysis of *DECLINE* which leads him to conclude that the different inflected forms of a lemma behave very differently from each other. Inflectional patterns are thus treated both as independent items and as parts of a lemma, where they are compared to the other forms. We assume that by default each form is unique in its behaviour, and try to find shared common features between the forms, rather than presupposing them.

Stubbs (2001, 64) lists four types of lexical relations from Sinclair (1991) and Sinclair (1996b); these partly correlate with Hoey's questions:

1. collocation

2. colligation

3. semantic preference

4. discourse prosody

Of these, *collocation* will be covered in the chapter on lexis, and that should answer Hoey's second question. The answer to question four, *colligation*, has been assigned to grammar rather than lexis, though it is on the border line between the two. We have already seen that *discourse prosody* (question three) is not suitable for automated analysis, because it requires value judgments, and *semantic preference* is a more restrictive variant of discourse prosody, which leaves aside questions of speaker attitude; but it requires assigning labels to groups of semantically related words. As with syntactic phrases, finding labels for semantic groupings is hard to do, and therefore we will not analyse semantic preferences here.

Collocation, on the other hand, is a central phenomenon in the description of lexis, and will be described in some detail. It is rather vaguely defined and thus it would be better to refer to it as a class of procedures that implement try to compute collocates using different algorithms to do so. There are many parameter values that need to be chosen which have a significant influence on the outcome; and hardly any standard or default values exist for most of these parameters. This lack of definition leads to great variability in the outcomes, so that results are rarely comparable.

We conclude the chapter on lexis with a discussion of multi-word units. These are arguably on the borderline between lexis and grammar, but since we are using a word form as the starting point for investigating its environment it would fit better into the area of lexis. There are several algorithms for identifying multi-word units, and several

of these will be investigated.

3.1.2 Grammar

The chapter on grammar begins with colligation, which builds on the results of the multi-word unit analysis from the lexis chapter. We interpret colligation as a more generalised form of multi-word units, where individual lexical items are replaced by category labels in order to reach a higher level of abstraction or applicability.

We then investigate grammatical relations between words. For this we need to fall back on the traditional categories of subject, verb, and object, as we will be doing a clause-level analysis of phrases that fulfil those roles, and we will look at which lexical items habitually co-occur for example with a given subject. The common relations in which a word is involved are called its *usage patterns*.

A further grammatical issue is that of 'pattern grammar' as described by Hunston and Francis (2000). Pattern grammar and the related 'local grammars' are in a strong position to become the main descriptive formalism of syntactic regularities in corpus-based grammar.

3.1.3 Meaning

Semantics has always been an area which many empiricists avoided, for reasons given by Sampson (2001). However, in recent years advances in corpus processing have made it feasible to venture into the realm of meaning. In this section we will look into approaches to find semantically related words. This will in part answer Hoey's first ques-

tion regarding the meaning of a word.

We will start off with a straightforward way of defining meaning: shared collocates. If a word shares many collocates with another one, then it is likely to share large aspects of its meaning as well.

Another, similar approach makes use of the usage patterns described in the chapter on grammar. If words share usage patterns, they would also share aspects on meaning, in the same way as it would work for collocates; only this time the relations are more well-defined, as they have a syntactic justification, whereas collocates are purely based on proximity in the text.

Continuing the theme of shared context we will conclude the analysis of meaning with an application of multi-word units. We treat them in analogy to the collocational frameworks of Renouf and Sinclair (1991), and argue that words which can substitute a given other word in its multi-word units have commonalities in meaning.

3.1.4 Multi-level Analysis

Several of these procedures described in detail in the following three chapters will take concordance lines as their input, and will classify those lines according to some feature that is contained in them, such as the tense or aspect of a node verb, or the grammar pattern of the node word. These classifying procedures are in the first place used as simple counters, which means they count and add up how often each recognised feature occurs in the set of lines.

The purpose of this first level of analysis is to describe features of the words in terms

of their frequency. By comparing the frequencies of the various features we can gain insight about how a word is most commonly used. We could for example find out that the verb *jilt* is never used in the third person singular, but much more frequently in the past tense or as a noun modifier. This is specific to the verb *jilt*, other verbs behave differently.

The second level of analysis would then be to describe co-occurring patterns. Apart from the straightforward frequency counts we can identify a number of combinations of features that are commonly used. For example, different tenses of a verb could require different grammar patterns, or they could be used with different collocates. By looking at either the tenses or the collocates independently we cannot find out those hidden regularities. So we create different sub-datasets which we investigate further, such as the concordance lines in the past tense being fed into the collocation procedure.

By doing this sort of analysis we can further restrict the usage potential of the word, provided we find any dominant patterns. We cannot predict the result in advance, and it will most certainly be different for different words.

The third level of analysis is then to look at the vocabulary in total and identify other words which exhibit a similar behaviour. Just as Hunston and Francis (2000) discovered that verbs with similar meanings make use of similar grammar patterns we can expect that words which have some features in common share other properties as well. This can be interpreted as increased redundancy, in that the same meaning is expressed both through the lexical choice and the grammatical pattern, and perhaps also through the choice of tense and collocate. Obviously this would not be an absolute, since the word meanings are not exactly the same, and words with different meanings also share the same (limited) set of grammar patterns. But all these features will contribute to the

overall meaning, and even if one was missing, the presence of the other features would compensate for that.

Another possible view is a reduction in cognitive load, though this has to remain speculative. Similar words sharing similar structures and features could indicate that they are stored in a mental lexicon where all those properties are combined, thus reducing the overall storage requirements.

3.2 Methods of Analysis

In the second chapter we have briefly discussed several previous attempts and approaches to a fully automatic and empirical analysis of language data. In this section we will investigate in more detail the methods used for the present study.

There are a number of issues in corpus processing which are relevant for the case studies mentioned above, which we will describe in detail in the following three chapters. Often discussion of such issues is ignored or deemed to be unimportant, but choices made at a rather low level of analysis filter through and can heavily influence the outcome in often unexpected ways. Especially in a study such as the present one, which tries to assess the feasibility of a certain methodology, one has to be open and transparent about everything in order to ensure replicability of results.

3.2.1 Units of Analysis

When investigating the structure of language one basic question concerns the units of analysis. There are several options when dealing with written data: letters, morphemes,

words, groups of words.

Letters are the smallest units available in written corpora. It might be possible to start off with identifying regularities on the basis of frequently recurring letter combinations; the basic discovery approach of segmentation and classification could lead to larger units, which on a linguistic basis might mirror morphemes as derived from phonemes. However, we would expect that the result would be overly influenced by the conventions of writing system used to represent words.

Morphemes, the basic elements in language that carry meaning, seem a more logical choice as the unit of analysis. In principle it should not matter whether a certain word form or one of its derived forms co-occurs with another word in question, so a collocational analysis would benefit from a morphological analysis.

However, Thurmair (1984, 177) states that *[t]he result of our research on morphemes tends to support those who stress the status of the word as the basic unit of linguistic theory*, as they cannot find a reliable correspondence between morphemes and meaning: *Words which are morphologically related [..] are completely different from a semantical point of view.* After analysing large amounts of data his research group at Siemens has not found any certain rules to derive the meaning of a word from the meaning of its constituent morphemes.

The word is the basic unit of lexicography. It is the basic index term, in that dictionaries (and even encyclopedias) are arranged in order, with a word as the key used for looking up information. Multi-word units are sometimes found under the entry of the main word form contained in it, e.g. *in spite of* can be found as the fourth sense of *spite* in Collins (1991). Paradigmatic variation in sentences is generally based on substitution

of words, and in syntax words are assigned to classes reflecting their various properties. Psychologically the word is the natural unit, and it is often used for tests such as word associations.

Multi-word units (MWU) comprise a variety of word sequences of different types. They include idioms, fixed phrases, and word sequences that behave like a single word (e.g. the preposition *in front of*). MWUs also seem to be the basic unit of meaning, providing context to single words that is required for determining its meaning. Single words out of context have multiple potential meanings and grammatical properties, but in actual usage in a text or as part of a larger unit this is general disambiguated.

However, MWUs are conceptually more complicated to deal with. Unlike single words they need to be identified in the text, and they could also have multiple 'phenotypes', where various surface realisations belong to the same conceptual MWU. Examples would be transposed word order, additional words that are inserted or left out, or morphological variations of the MWU's elements. At this stage the properties of multi-word units are not known well enough to feel confident with using them as the basic unit of analysis.

In the present study we will therefore use the word as the smallest unit of analysis. There are clearly problems with this, especially when looking at meaning, but we have to accept that we cannot solve all problems at once. Ideally we would want to use words and multi-word units where appropriate, but allowing for combinations of words introduces an additional layer of complexity that might have obscured the basic argument this thesis tries to support.

3.2.1.1 Orthography and Tokenisation

The basic element of corpus processing is the *token*, usually based on an orthographic sequence of characters not interrupted by white space or punctuation. The inverted index of the corpora used contains tokens, so in order to find any linguistic element without sequentially perusing the complete corpus we need to know how it relates to tokens. In a software demonstration of the Qwick corpus browser at a conference the author was quite surprised that the BNC sampler did not contain a single instance of *gonna*, until a member of the audience pointed out that in the BNC it would have been tokenised as *gon* plus *na*. The same applies to frequency lists, so it can be a non-trivial task to find out the frequency of occurrence of a non-obvious linguistic element.

As Leech (1997) remarks, there are three main cases where there is a serious difference between 'orthographic word' and 'morphosyntactic word': multi-word units (e.g. *in spite of*, or *New York*), mergers (e.g. *don't* and French *t'aime*), and compounds (e.g. *word class*, *word-class*, or *wordclass*). Difficulties in making decisions automatically lead to 'phantom words' such as *don* (as in *don't*, which was always split this way by the Bank of English tokeniser) or *York-San* (from the sequence *New York-San Francisco flights*). Treating all cases the same leads to a certain class of errors, but that is unavoidable with fully automated processing.

Sampson (1995) suggests specific rules to deal with such cases, but they would require manual processing, which would be unfeasible with corpora beyond a certain size.

It might be useful to compile a list of multi word units in advance, so that the tokenisation can take those into account. While this would solve the *don't* and the *York-San*

problem, some cases require different kinds of knowledge, as for example the multi word conjunction *provided that*. One can think of (admittedly invented examples) of the type *X provided that service*, where a simple pattern matching approach to multi word unit recognition would fail. However, this is likely to be extremely rare, as almost 1400 occurrences of *provided* in the BBC corpus contain not a single instance of such a construction. An analysis of the written part of the BNC (more than 16,000 occurrences of *provided*), though, quickly shows up a few counter-examples (*provided that necessity* and *provided that boat was big enough*, where *provided* alone is the conjunction). This again shows that linguistic judgments are often sensitive to the text type under investigation and any presuppositions based on intuition should be avoided.

Case is another difficult issue. The software for accessing the Bank of English makes no distinction at all between upper and lower case variants, which simplifies retrieval, but favours recall over precision. Maintaining a distinction between different case variants not only increases the physical size of the inverted index, but also makes processing more difficult, and some systematic errors might occur if certain words occur predominantly in sentence-initial position.

It is hard to make a decision either way. By working non-case sensitive some important distinctions might get lost, and names like *Brown* might interfere with the collocational patterns of 'ordinary' words. Keeping up the distinction hurts recall, as fewer relevant forms are found, but maintains precision. At this stage of research we would rather sacrifice some of the comprehensiveness of coverage than blur the overall results. For that reason we have decided to treat upper and lower case variants as different types. As a consequence some patterns we discover might in reality be more pronounced, as we are missing out the variations with different case. On the other hand, we might not find some patterns which are deemed not to be frequent enough if the proportions of

upper and lower case versions are similar in size.

3.2.1.2 Lemma, Lexeme, Lexical Item

Knowles and Don (2004) discuss the problems caused by the 'English' definition of lemma, and state that it is a fairly useless concept, especially when compared with other languages such as Arabic and Malay. Languages with richer morphology structure their vocabulary in very different ways from English, so that for example *singer* would be part of the lemma *sing*. Traditionally these would be two different lemmas in English, though *singer* is linked through derivation to *sing*.

Their definition of lemma is as a set of variant forms:

$$DEAL = deal, deals, dealing, dealt \tag{1}$$

It is an open question whether that should include word class variants, such as deal_VB and deal_NN. Also, *deal* and DEAL are two different conceptual entities, the former being the graphic form (base form of the verb *to deal* or the noun *deal*), the latter one referring to the lemma as defined above.

The specification of 'variant forms' is made more formally explicit by Allén (1981, 382), who quotes his own definition of 'lemma' from 1970:

> A lemma is defined as a group of forms within a word-class which are assignable either to one and the same series of inflection (in the case of indeclinable words comprising only one form) or to several series of inflection that converge in speech and/or in writing, the divergences of which

show purely facultative variation (free variation).

He also introduces the concept of *lexeme*, which is basically equivalent to a word sense. He gives the example of Swedish *sticka*, which belongs to three lemmas (one noun and two verbs) with a combined total of six lexemes.

Crystal (1992) simply defines 'lexeme' as another word for 'lexical item'; as a lexeme subsumes inflected forms it is equivalent to Allén's 'lemma'. However, in the remainder of this text we will use Allén's meaning of 'lemma' and will avoid the term 'lexeme' altogether, speaking instead of 'inflected form' or simply 'word form'.

The basic question relevant for this project is that of the unit of analysis: the lemma is an abstract concept, even though a discovery procedure could probably identify a link between morphologically related words as long as they are regular (e.g. *talk* and *talk-ing*). The first problem concerns orthographical adjustments (*swim* and *swimming*) and then irregular forms (*go* and *went*). A possible solution might here be that of semantic similarity: if we investigate usage we might find that *go* and *went* are used in the same contexts and could thus be identified as related.

In fact, a semantic analysis of *went* (using a procedure detailed below, see section 6) shows *go* and *gone* as the two most similar items based on the verb-object relation. Despite morphological differences we cannot rule out the possibility of an enhanced discovery procedure making use of both local, low level character distribution and global, syntactic, phrase-level distribution.

For the time being we will work with individual word forms, defined as strings of characters without intervening spaces or punctuation other than hyphens or apostro-

phes (which receive special treatment). The definition depends on the tokeniser used for segmenting the corpus. A lemmatiser (currently only for English language corpora) will be used to group word forms together, mainly for presentational purposes, but also for certain types of analysis, such as frequency distribution of inflected forms, and shared collocations. Other than that, each word form is investigated independently of all others.

No attempt is currently made to resolve homographs, a problem that has been identified early on in automated corpus studies (Sinclair *et al.*, 2004)

3.2.1.3 Word Classes

Word classes are an attempt at syntactic generalisation. Their purpose is to assign words to classes which exhibit similar syntactic behaviour. The idea behind this is that the grammar of a language can be described with rules using a small amount of classes instead of the actual words. However, as described in section 2.3.2 above, most word class systems are ultimately based on a view of grammar derived from Latin and Greek during the Middle Ages, and consequently do not suit other languages very well.

For the current project word classes are used for the identification of the properties of verb groups (section 4.1.4) and to facilitate more sophisticated grammatical analysis: for colligation (section 5.2) they are used to capture basic abstractions, with usage patterns (section 5.3) they are used for a shallow sentence analysis with the aim of extraction sentence roles, and for grammar patterns (section 5.4) they are used to identify phrases with a parser. For all of those procedures it is necessary to have a grammatical description in the form of (a few) rules, so that phrasal units can be identified.

3.2.2 Corpus Data

3.2.2.1 Representativity

Many linguists consider the *representative corpus* as the holy grail of corpus building. This (hypothetical) corpus contains the best possible balance between different text types or genres, and avoids any bias towards a specific kind of material. Tognini-Bonelli (2001, 58) states that *ideally, a corpus should be unassailably representative*. However, we can consider this ideal as both unobtainable and undesirable, for several reasons:

First, language is not at all homogeneous, and too little is still known about the exact differences between its different manifestations. In actual fact, if we view a language community as a cluster of people with near identical idiolects (however we want to operationalise 'near identical'), then a language could be the sum/average/common denominator of the idiolects of the cluster members, clearly not a well defined entity However, this model would provide us with a very neat definition of the otherwise purely fictitious 'native speaker', which would be the medoid of such a cluster, the central element, which could be interpreted as a prototypical instance.

Second, no single speaker of a language is exposed to the same kind of utterances as everybody else, so a representative corpus would require some kind of abstract 'native listener' whose experience of language would be modelled; this also introduces the problem of production *vs* perception: a corpus probably ought to represent the latter, whereas some corpora are concerned with the former (e.g. the corpus of Dickens' work, used in the present research).

And third, few investigations would actually be interested in such a corpus, as they are (at present) more concerned with the properties of certain subsets of language, as their respective properties can be described more easily. Though not restricted to any particular kind of data in theory, in practice most research is based on a limited amount of data, either for opportunistic reasons of availability, or because the data under investigation has certain properties which make processing easier (e.g. when using written rather than spoken material).

Monaghan (1979, 50) summarises the need for selecting a sample of text for analysis: *Without some sort of selection it is impossible to make statements which are both meaningful and applicable.* One can either idealise and describe language free from contextual or situational 'distortions', in the way Chomsky has done with the concentration on *competence*, or one can limit the scope of the description to a small and well-defined subset of language that can be analysed reasonable comprehensively, such as the language used by Dickens for writing his novels. As Monaghan points out, these two scenarios lie at the extreme ends of a continuous scale. For the same reason, Kittredge (1981) advocates the study of sublanguages, which are less complex and more limited in their use of grammar and lexis than unrestricted language. A sublanguage is by definition not representative.

Rieger (1979) describes in detail why he regards *representativity* as an inadequate concept when dealing with corpus data: in order to identify a representative subset we need to know the whole, the overall population. But he likens corpus building to random sampling from an *unknown* population, which requires statistical procedures for reaching valid conclusions. Unless, of course, we know the complete population, as we would when composing a representative sample of Dickensian novels.

We can further clarify the difference between a representative sample and a random sample by an example: before an election we can pick a representative constituency whose result in the *previous* election was most similar to the overall outcome. This can now be used to predict the next election on polling day, and instead of having to wait for all votes to be counted accurate predictions can now be made quite early on by just evaluating a few key constituencies. Afterwards the prediction can be compared with the end result and we can see whether the chosen constituencies were really representative of the whole electorate. The important issues here are that we know the *Grundgesamtheit* (sampling space), the final result of the previous election, and we use that knowledge to select our representative sample.

But in language we do not have this information available. Any corpus is simply a random sample from an unknown whole, so calling a corpus *representative* is simply inadequate and wrong. The term *balanced* which is often used simply complicates matters by mixing different samples together in (usually) arbitrarily predetermined proportions to avoid the bias of having just one text type or genre available.

However, the best kind of corpus for most studies is a *homogeneous* one: it allows to state exactly for what type of language your conclusions are valid. Obviously the results are more generally valid if they can be repeated with other data as well, but then it is better to work with distinct data sets rather than a mixture. It is always possible to combine several such corpora to form a working set comprising different text types. As corpora (or rather, the software used to access them) typically lack facilities for detailed selection of subsets, it is better to be able to combine different distinct and homogeneous sets of data.

This, of course, then brings up the question of *homogeneity*. Any text has a large

number of features which can be considered when assigning it to a homogeneous sample: date, genre, register, authorship, source, *et cetera*. Deciding upon a single criterion is clearly arbitrary, but at least it can be objective, and other researchers could change it when repeating the work to see whether it had an actual influence on the outcome.

Any results of corpus analysis are necessarily restricted in their validity. A frequency list of the Times newspaper will not allow one to make statements about the frequencies of words in novels by Douglas Coupland. Collocations of `ship` from Douglas Adams' *Hitchhiker's Guide* books will allow no predictions of collocations of the same word in C.S. Forrester's *Horatio Hornblower* series. However, that is not unusual: no geologist would think of looking at the layers of rock on the Isle of Wight and making any assumptions about rock formations in Alaska or even on Mars.

It is of course possible to make such predictions, but they are subject to testing with the appropriate data. In the course of scientific discovery this is a perfectly natural process: predictions which hold for more and more data will grow stronger over time, whereas weak predictions can be disproven quite quickly on a different corpus. Obviously, the safest prediction will be one limited to a narrow range of data: it will be harder to find appropriate data that fits into this range where the prediction does not apply. However, this means that the prediction will not tell us very much about language.

Here we have a trade-off between the strength of claims made and their scope. Typically, the broader the scope, the weaker the claims, as more possibilities for counter-examples exist. Strong claims can only be made for small and restricted data sets such as sub-genres. Tognini-Bonelli (2001, 59) recognises that issue in her section on sampling, when she advocates that the corpus creation criteria should always be made explicit,

so that users are *in a position to evaluate the corpus using the criteria and relate the statements they derive from the analysis of the corpus to the typology of the texts included in it*.

3.2.2.2 Subcorpora

For the present research a number of different corpora have been collected. As the aim is to investigate methodology, rather than analyse a particular set of data, the collection has been purely opportunistic. The guiding principles were to get some reasonably large amounts of more or less homogeneous data, but also to sample a wide variety of data. No claims are made on the distribution of these samples within the total of language; therefore all outcomes are necessarily hedged by the caveat that they might only be valid for those particular corpora. However, in principle comparable results should be achievable with other material.

This 'opportunistic' corpus building provides us with an additional advantage, namely that we can evaluate the results more easily. Using a corpus restricted to a particular subject area reduces the variability of word uses. And using obviously different corpora (such as newswire articles and nineteenth century novels) we can make sure that the outcome reflects the language used in the data and brings out those differences.

The following corpora were used during the development phase of this thesis, with a combined total of 166 million tokens.

blt Birmingham-Lancaster corpus, set up for a project between the two universities on tagger evaluation. About 1.2 million tokens of general texts.

lob The LOB corpus.

flob The Freiburg LOB corpus, a 1990s update of LOB.

frown The Freiburg Brown corpus, a 1990s update of Brown.

mcon A sample of about 200,000 tokens from the MicroConcord text collection, journalistic texts.

BBC The Bank of English BBC World Service subcorpus

bncW0-9 The written part of the BNC (split into 10 parts).

bncSpok The spoken part of the BNC.

19C A collection of 19th Century novels.

3.2.3 Software

This section contains a description of the general software developed for the analysis. In general we can distinguish two major types of software: those used for gathering data, and those for processing it. Some types of data require multiple processing steps, so that sometimes the distinction between collecting and analysing is not obvious. We will be talking here only about the basic software that is used throughout this project. The specific software used for a particular research task is described in the appropriate section in the following three chapters.

There is a large amount of processing software available; some programs have limits as to the amount of data they can process or possible processing steps. While many of them are well suited for corpus analysis, few of them are flexible and 'open' enough for

more fundamental research. Many procedures involve a number of parameters which can be varied, but often programs use default values for those which are not shown to the user and cannot be altered. For a more detailed discussion of this issue see Mason (2000a).

This excludes much corpus processing software from use for the research described in this thesis, as the systematic variation of parameter settings is important to explore what their respective effects are on the outcome. For this reason, and to have a coherent integrated system, all software used in this corpus has been developed separately, though not necessarily in connection with this research. A lot of the software was written before, see e.g. Mason (1996) on a description of the principal way the corpus access was implemented.

The following section lists a number of procedures used for this research and describes the tools used to apply them to the data. Without exception the software is implemented in the programming language Java, which is very suitable for this kind of processing and facilitates re-use of individual components. For an introduction to using Java in corpus linguistics see Mason (2000c)

CORPUS ACCESS was implemented in a form that provides an API for all fundamental functions: retrieving word frequencies, a word list, and concordance lines. These lines can later be processed to provide contextual information of words. The corpus data is kept in plain text format, but is tokenised while a word list is created. All tokens are stored in an inverted index for fast access. The indexing procedure has been implemented after Witten *et al.* (1994), and had already been tried and tested in the CUE corpus access system (Mason, 1996) on which the current system is partly based.

STATISTICAL ANALYSIS includes a number of univariate and multivariate procedures (Correspondence Analysis (Greenacre, 1993), Principal Components Analysis) as well as Cluster Analysis (PAM and AGNES, see Kaufman and Rousseeuw (1990)). All routines were coded in Java to allow for seamless integration with the other modules. The statistical procedures are described below in more detail.

LINGUISTIC PROCEDURES used for this thesis include lemmatisation and parts-of-speech tagging. A lemmatiser for English used is based on a description in Harris (1985). Tagging has been performed with QTag, a POS-tagger based on stochastic principles (Tufis and Mason, 1998).

Danielsson (2001, 101) questions the use of linguistic procedures for the automatic discovery of linguistic units. She is of course right to point out the interference of the paradigmatic view on the results, so we need to be careful to choose whether it is worth using tagging and lemmatisation whenever the question arises. If we postulate that language behaves according to certain universal regularities (such as Zipf's Law, and general principles gained from information theory), we can perform the analysis multiple times with different settings and decide objectively which of these combinations yields the best result. It is perfectly valid to state that each inflected form of a lemma behaves in a completely different way with respect to linguistic features under investigation, but there is no principal reason why this has to be the case. Stubbs (2001) gives examples of *commit*, where collocations are those of the lemma, and *seek*, where collocations are restricted to a particular inflected form. If all inflected forms (or all but one) exhibit the same behaviour, then lemmatisation provides us with a useful generalisation over the behaviour of a number of different word forms. The important point here is not to have preconceptions either way which would bias the outcome.

3.2.4 Statistical Procedures

The type of analysis used here is *inductive*. We start with the data, and try to discover patterns in the data using statistical methods. This contrasts with the *deductive* approach, where one would form a hypothesis and then try to falsify or prove it by analysing the data. But we do not want to bias the outcome of the study by introducing preconceptions about language, so the deductive approach is not suitable for the current project.

In statistics the kind of inductive data analysis is called *exploratory data analysis*, as we effectively explore the data while looking for significant patterns. We obviously need to have an idea what those patterns could look like, but this can be stated in very general terms: a) entities which share features/properties can be grouped together, and b) features which are shared by similar entities can be combined. The two statistical methods we will be using for these two kinds of patterns are *cluster analysis* and *correspondence analysis*.

3.2.4.1 Cluster Analysis

Cluster analysis takes as input a set of entities described by feature vectors, and attempts to group them together using a similarity metric that for two feature vectors computes a single value representing the similarity (or dis-similarity) of the related object. Depending on the clustering method used the most similar pair of the data set is joined up either in a lump (agglomerative clustering) or inserted into a tree structure (hierarchical clustering). One problem with cluster analysis is that it will always find a structure, even if there is none present in the data (or if it is completely random). For that reason

it is useful to be able to evaluate the quality of the result; and there is a quality indicator available for the agglomerative clustering method used here (Kaufman and Rousseeuw, 1990). For hierarchical clustering a visual inspection can usually determine whether the result makes sense.

With agglomerative clustering one usually has to specify in advance how many clusters the system should find. This poses the potentially serious problem that we have to define in advance how many groups there have to be, and the algorithm then only partitions the data into that number of groups. However, using the quality metric we can easily solve that problem by iterating (automatically) over a number of different cluster numbers, and choosing the result which has the highest quality value at the end.

The most common similarity metric in cluster analysis is the Euclidean distance, and even though there are a number of alternative metrics available there is no compelling reason not to use it, as long as the feature vectors contain numerical values. A different metric would have to be used for boolean values. Oakes (1998) lists a number of available metrics.

3.2.4.2 Correspondence Analysis

Correspondence analysis is a type of factor analysis, especially geared for the analysis of contingency tables (Greenacre, 1993). It compares the value profiles within both rows and columns to identify the contributions of rows and columns to the overall variability, and allows projection of the table onto a two-dimensional plane; this is effectively equivalent to multi-dimensional scaling. Both row and column labels can be projected on the same plane (though with differently scaled axes), which allows for easy interpretation of the results.

3.2.5 Data Processing

The choice of data formats is important for the work described in this part. Various different kinds of information are gathered, and some data will be immediately re-used for other processing steps. At the end of the processing, all relevant data will be presented to a human user, though it should be possible to use the data from within other computer applications. The data format therefore has to fulfil two major requirements: it has to be *flexible* enough to allow storage of various types of data, and it has to be *accessible* by both human beings and computer programs. The latter are more important, as the 'raw' data will only be inspected during the research and development stages, whereas end-users would normally have the data presented in a more ergonomic way by the computer.

Binary formats are compact and can easily be written and read by computer programs, but are not very flexible, and, most importantly, are not portable in the sense that it is hard for other applications to read them. It is also not possible for a human to read them without the aid of special software, and even then it requires complicated deciphering of the numerical codes. There could also be problems with byte order across different computing platforms, as processor architectures use different ordering conventions when storing multi-byte sequences (for example for storing integer values larger than 255).

Plain text formats use text to encode the data, in a kind of explicit 'long' form: numbers are represented as sequences of characters representing digits rather than internal binary values. This makes the data more verbose, but on the other hand it is now a lot easier to read this data as a human being. There are standardised functions

for computers to re-interpret those textual representations and to convert them back into numerical data. However, most plain text formats have to perform a balancing act between two extremes: ease of use by humans versus ease of use by computer.

A very restricted format is easier to process by computer, but might be more cryptic to read for a human; a more verbose format adds additional information to ease analysis by a human, but could then make it harder to read in by computer. An ideal solution seems to be a class of plain text files which has been standardised: XML.

XML (eXtensible Mark-up Language) is a verbose format that is clearly specified, and can be easily processed by computer. A large number of parsers for XML data exists in a variety of programming languages. These parsers differ in the extent to which they require formal correctness of the input data, and also in the paradigm used for dealing with the data. Two different approaches are used for processing XML data: DOM (document object model) where the input data is read into a tree structure in memory, and SAX (Simple API for XML) where a document handler module is defined that provides callback which are called whenever a certain event occurs, such as the reading of an opening tag, a closing tag, or the end of the document. The document handler can then process the data as required, even if not all the input data is available. In DOM the data processing takes place through tree-manipulating methods, but this requires that all the data is available.

For the current project we use individual files to store information for each word, which means that the file sizes are not very big. For that reason it is easier to use a DOM-style processing model. The software package XOM (Harold, 2002) was chosen for ease of use, being fully compliant to the standard, and efficiency.

In order to use XML to store data it is necessary to define a schema/DTD that describes how the data is mapped onto the tagged mark-up format. Technically this is optional for XML, but is required for practical reasons, as the document processing module needs to be aware of the possible range of tags. Again, there are two options:

- Firstly, a specialised schema can be created for each data type. This has the advantage that it can be fitted perfectly to the shape of the data, and it can also be used to validate the integrity of the data as it is created or read in. The drawbacks are that it is inflexible, ie it needs to be changed every time the specifications of the data change, and it requires specific document handler modules that need to implemented for each type of data.

- Secondly, a generalised schema can be designed which basically works as a broad skeleton and provides a rough structure that can be filled with the data on an individual basis. This has the advantage of flexibility, requires only a single document handler, and can be implemented as a general utility module, a sort of sophisticated container for general data. The drawbacks are that being completely general the data cannot be validated, and the created data has to be squeezed into a frame that might not be a perfect fit.

Both options suffer from the unavoidable coupling between different modules, as any module reading the data at a later stage needs to be aware of its specification, and is thus vulnerable to change. It might require less effort to use a broader schema, as the document handler remains constant, but a disadvantage is that it might not be as easy to detect incompatible changes that have been made through software revisions, due to the lack of a validation facility.

The option chosen for the system described here is the first one. Data is generally stored as individual entries with attribute names encoded in the tag labels, and attribute values as textual data, but more complex constructions are also possible. An example of the output produced by the system is given in the appendix B.

One problem that a verbose format such as XML creates for storing large amounts of data is the required space. Binary formats are more space-efficient, though this is not a real problem with increasing availability of storage media. However, modern programming libraries allow easy application of data compression methods, which can be used to keep the size of the data files small should size become an issue.

3.3 Summary

In this chapter we have discussed what areas of language we are going to investigate. Due to space and time constraints there has to be a limit as to what can be covered. One aim of this selection process was to have a spread over the three main linguistic disciplines of lexis, grammar, and meaning.

We have also described the methods of analysis. This includes the data to be analysed, the units we are looking for in that data, and the computer software and data formats used for processing. We have also mentioned the several levels of analysis, some of which are only made possible through automatic analysis.

Chapter 4

Lexis

In this chapter we will discuss information relating to lexis that can be gathered automatically. There are three parts to it. The first part covers the distribution of word-related features, such as frequency, variation based on inflection and derivation, and also distribution of number, tense, and aspect. For the second part we move from the intrinsic features of a word to its environment, looking at words which frequently co-occur. There has been a lot of research in the area of collocation over the past few decades, and we will here try to distill some of the information that can be gained from the analysis of word co-occurrences. And finally, the third part moves towards the next chapter, grammar. Multi-word units are located somewhat between lexis and grammar, and might thus warrant their own chapter. However, the approaches discussed here take the (single) word as their starting point, and so it was felt most appropriate to include them in the chapter on lexis.

4.1 Lexical Statistics

Some fundamental statistics that are generally informative about a word's nature can easily be collected. These include the general frequency of occurrence, morphological variations, and phrasal features such as tense and aspect of verbs.

Most of these statistics are not meaningful on their own, but they become meaningful when contrasted with others. There are multiple ways in which this can be done: different words can be compared with each other, or the same word could be analysed across a set of different corpora. Or even the combination of the two, where the differences among a set of words are tested across a variety of corpora to see whether their relationships are stable and independent of the data source or not.

In this section we will discuss these basic statistics in more detail, before we continue with the more advanced processing steps in the next sections.

4.1.1 Frequency of Occurrence

The *frequency of occurrence* reflects the relevance of a word in the trivial sense that higher frequency words are generally frequently encountered. It is given as an absolute value, in order to give an idea of the reliability of subsequent statistics (as described in the following sections). For cross-corpus comparability we also state the frequency per 10,000 words, and each word is is also assigned to a *frequency band* (Quasthoff, 1998). The frequency band of a word is derived by dividing the frequency of the most frequent word in the corpus by the frequency of the word in question, and taking the base-2 logarithm. The resulting frequency band is in the range of zero to about twenty,

where zero means it is as frequent as the most frequent word (*the* in a non-exceptional English corpus). The largest frequency band contains the *hapax legomena*, which have a frequency of 1 (Quasthoff's range extends to 21, which is based on the size of his corpus). Words not occurring in the corpus would be assigned a frequency band of -1. The scale is logarithmic, thus taking into account the power-law nature of Zipf's Law (Zipf, 1935).

The total distribution of frequency bands in the written part of the BNC is shown in figure 4.1.

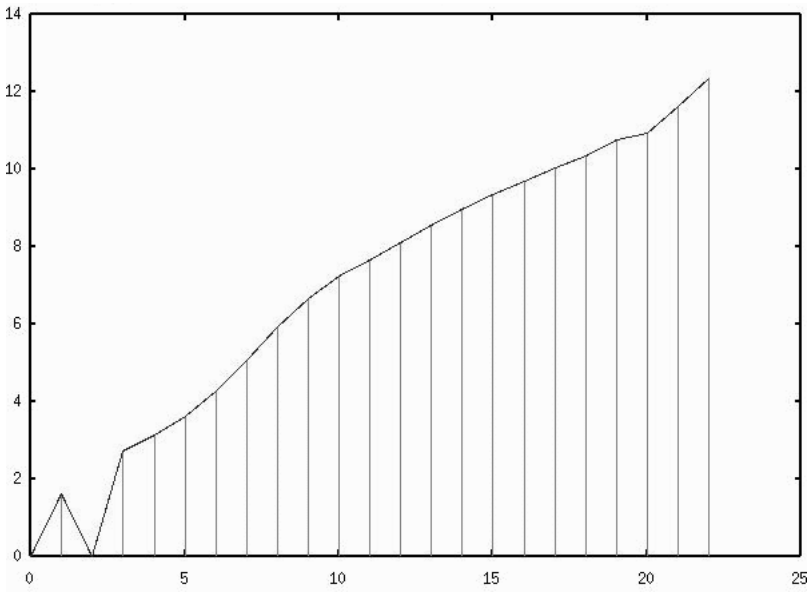

Figure 4.1: Distribution of frequency bands in the written part of the BNC corpus

The vertical axis represents the logarithm of the number of word types contained in the respective frequency band. Without taking the logarithm the curve would be close to the horizontal axis for most of the bands and would then rise sharply towards the end. With a corpus of roughly 90 million tokens the number of frequency bands is 22.

The distribution of bands in about 360 million tokens of mixed corpora is as shown in figure 4.2 (roughly 960,000 types).

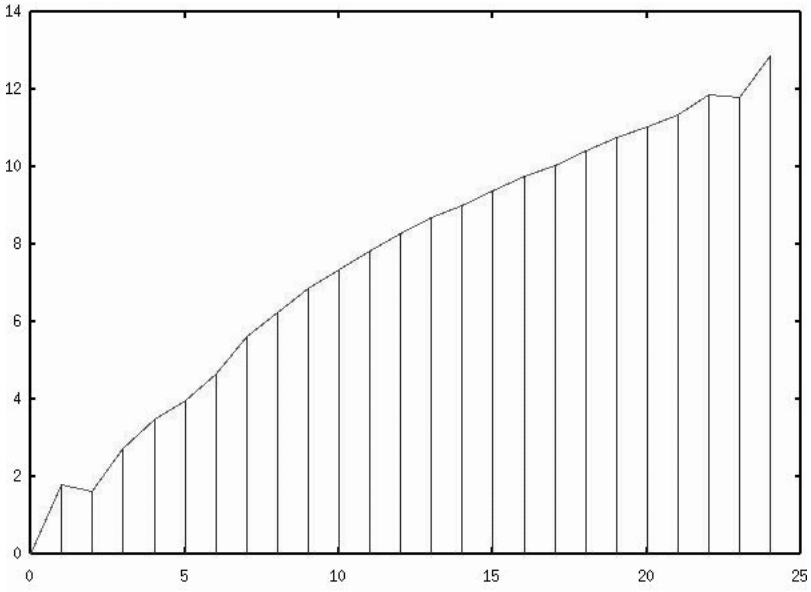

Figure 4.2: Distribution of frequency bands in 360 million tokens of mixed corpora

As we can see, they are mostly identical, apart from the high-frequency words, which interestingly seem to vary considerably in their distribution. However, this is a common problem with logarithmic scales, as small counts are represented disproportionately: the first four values for the written component of the BNC are 1, 5, 1, and 15, whereas they are 1, 6, 5, and 15 for the complete set (which includes the BNC). Also, the number of frequency bands is now 24, which reflects the fact that the corpus is four times as large as the (written) BNC.

From the analysis of this distribution we can conclude that the frequency band parameter provides a convenient way of abstracting away from the actual frequency values which are subject to minor variations due to chance. By 'smoothing' the values in the described way further processing will be more robust with respect to those random

fluctuations.

The frequency of occurrence is directly accessible from the corpus access software, as it will have been used for the creation of the inverted index as well as for the retrieval of instances. The frequency band is computed from the frequency of occurrence of *the* compared to that of the word in question. Furthermore the number of occurrences per 10,000 words is calculated. If the corpus consists of more than one subcorpus, the distribution of frequency bands across subcorpora is also calculated, in order to reflect the textual specificity of the word.

4.1.2 Distribution and Spread

A further parameter that describes the overall behaviour of a word is its *spread*. The spread of a word indicates whether its instances are evenly spread across a corpus or occur in clusters. This is expressed through a value between 1.0 (maximally spread out) and 0.0 (maximally bunched). The value is calculated by dividing the corpus into a number of evenly spaced compartments, one for each occurrence of the word. Then all occurrences are distributed across the compartments, and the number of non-empty compartments is divided by the word's frequency to yield the spread coefficient.

The spread value is not independent from the word's frequency. As the number of filled compartments is divided by their number, high-frequency words have a larger divisor than low-frequency ones, which makes their spread values not easily comparable. Furthermore, high-frequency words have smaller compartment sizes, which makes it more likely for random variation to affect the distribution of word forms into compartments, whereas this is not as much of an issue for low-frequency words with larger

compartment sizes.

For this reason spread values of words with large frequency differentials cannot directly be compared, and word frequency has to be taken into account for the interpretation of the results.

Spread is further measured by the mean distance between occurrences, and enhanced through the standard deviation and variance. For technical reasons the values are computed in the number of characters rather than tokens, as the corpus access system used indexes the text data on character positions. However, this will have little influence on the values involved, apart from the fact that they will be larger.

A further coefficient describing the distribution of a word throughout the corpus is *coverage*. Here the corpus is divided into a fixed number of compartments, and then the value is computed the same way as the spread. The difference between the two is that the number and size of the compartments is equal for each word, irrespective of their frequency. The coverage thus describes more adequately whether a word is used throughout the corpus, and frequent words will almost automatically have a larger coverage than infrequent ones.

The compartment size is determined automatically, and it has to be in some relation to size of the corpus in order to be universally applicable. For that reason, the square root of the corpus size has been chosen. This does in fact mean that the number of compartments is approximately the same as their size in character positions (since the number of compartments is the square root of the number of characters in the corpus); only rounding errors introduced by the discrepancy between integer and floating point numbers involved in the calculation cause a slight differential.

Further values computed are the smallest and largest distance between occurrences.

Tables 4.1 and 4.2 give a few example values:

WORD	FREQ	SMALLEST	LARGEST	COVERAGE	SPREAD	MEAN	STDDEV
ball	285	10	2723014	0.0191	0.4947	381261.58	551230.80
cloth	28	35	10783431	0.0023	0.6071	3669267.96	3613647.46
coach	304	16	2940157	0.0193	0.4703	355273.72	550051.25
correspondent	17873	19	98505	0.6900	0.5264	6117.71	8759.49
dish	64	29	8755857	0.0049	0.5625	1625326.89	1809787.54
election	6505	13	464686	0.2911	0.4089	16800.65	32670.75
house	2359	9	653822	0.1416	0.4705	46327.80	68907.78
houses	1070	25	1141097	0.0802	0.5289	101695.00	133502.58
news	8492	13	173901	0.4688	0.5395	12875.44	17015.27
of	584074	3	14342	0.9999	0.6520	187.21	184.25
said	76820	5	62001	0.9484	0.4998	1423.44	2474.76
the	1184077	4	14543	1.0000	0.6721	92.34	88.87
voters	1252	15	1505683	0.0795	0.4345	87205.89	163885.45
votes	1504	10	1539274	0.0975	0.4521	72638.82	131430.43
war	10906	7	230898	0.4902	0.4792	10022.44	15493.84

Table 4.1: A sample of words from the BBC corpus

WORD	FREQ	SMALLEST	LARGEST	COVERAGE	SPREAD	MEAN	STDDEV
ball	6428	8	16371990	0.1051	0.2559	83169.61	328051.14
cloth	1915	24	16888043	0.0527	0.3697	279238.64	705916.98
coach	3072	8	11852704	0.0646	0.2490	174085.59	599049.20
correspondent	629	55	21192300	0.0222	0.4419	849381.00	1658617.84
dish	1464	8	10853811	0.0341	0.3203	364513.39	870925.43
election	9362	9	6996962	0.1287	0.2153	57027.98	260074.73
house	32309	6	2608840	0.4441	0.3706	16552.19	46693.21
houses	8567	8	3475417	0.1859	0.3763	62404.05	139817.85
news	10014	6	3262254	0.2425	0.4216	53394.62	124441.71
of	2912506	3	15047	1.0000	0.6296	183.62	204.27
said	180688	5	1167156	0.7317	0.3537	2959.63	12780.76
the	5143707	4	22475	1.0000	0.6393	103.97	114.41
voters	1803	12	22171025	0.0375	0.2113	296179.18	1143235.84
votes	2995	9	9620464	0.0560	0.2123	177400.19	665704.36
war	21517	4	3576559	0.2980	0.3109	24847.58	82426.44

Table 4.2: A sample of words from the BNC corpus (written components)

As can be seen from the table, *the* has maximum coverage with a value of 1.0, which is to be expected. The spread values do not seem to be as reliable an indicator of the way a word is used throughout the corpus, which indicates that there is a lot of local fluctuation, even with words like *the*. In fact, their high frequency means that

the compartments are smaller and therefore local variability has an undue influence. The mean distance between occurrences should probably be normalised by frequency to give a meaningful result, since the two values (mean and frequency) are correlated.

In the BNC sample, *cloth* and *votes* have very similar coverage values, but *votes* has about half as many occurrences as cloth, which means that it is more 'bunched', as those additional 1000 instances occur in compartments that are already occupied with previous occurrences. In other words, *cloth* is more spread out in comparison.

4.1.3 Inflection and Derivation

For our analysis we will treat individual forms of a lemma as separate items. As Stubbs (1993, 17) concludes from research in lexico-grammar, *[m]eaning is not constant across the inflected forms of a lemma*. So while we regard the inflected forms as elements belonging to a canonical form or lemma (following Allén 1981), they are treated independently as items in their own right, without any presuppositions about their distributional or other behaviour. The more detailed study of distribution of word forms across word classes is discussed in the next section.

Sinclair (1991) conducts a detailed study of the lemma DECLINE, in which he splits the frequency counts of the various forms according to the word class to which the instances belong and later also divides them according to sense. He concludes *that grammatical and lexical distinctions may be closer together than is normally allowed* (1991, 51).

While we can easily automate the first part of the study, so far we cannot do so for the second part, as there is no automated way of assigning word senses to instances

of word forms (and Sampson (2001, 200) reports work by Kilgarriff which casts doubt on the very existence of clear-cut word sense distinctions). The purpose of an analysis of the distribution of inflected forms across word classes is not so much to discover information about a particular lemma, as we cannot yet interpret the findings in the light of an appropriate theory, but rather to identify general patterns of language. By collecting statistics of word form distribution for all words studied we can try to identify tendencies, since we would assume that the distribution was not entirely random.

Unlike the other processing steps this procedure deals with one lemma at a time. All inflected variants are generated, and concordances are retrieved. Then we assign word class labels using a parts-of-speech tagger (Tufis and Mason, 1998), count how often each form occurs with each word class, and tabulate the result.

A further processing step which is performed at this stage is the analysis of tenses for verb forms. Using a simple finite state automaton the verb phrase is evaluated for tense and mood. Distribution figures for all possible results are added up and stored. This process is described in more detail in the following section.

4.1.4 Number and Tense/Aspect

The English language has a very limited inflectional morphology, and few forms are marked for grammatical features. For nouns this is the singular/plural distinction, and for verbs it is tense, where the morphological changes are often combined with further modals or auxiliaries to express the overall tense of the verb phrase they occur in. There are a number of purposes for which this data can be used:

- distributional patterns can be identified. It might be possible that verbs sharing similar tense distribution also share other properties. We can also find out whether there is any bias in the distribution across corpora or text types.

- knowing the frequency of occurrence of verb tenses is useful for the creation of teaching materials, where the most common tense forms of a verb can be introduced first. It might also give us more of an idea how a verb is typically used.

- differentiation: at a later stage we could look again at collocations (and other parameters) and distinguish them according to which tense they co-occur with. This would open another dimension to the description of a verb form that has previously been ignored.

- general tendencies can be analysed. One could assume that complex tenses were less frequent than simple tenses, or that certain tenses were predominant in certain text types.

Some words are notably biased in their distribution of these forms, and with others there is a strong correlation between inflected form and word sense. For example, in the 400 million words of the Bank of English there is not a single occurrence of *jilt* in the present tense, and Sinclair (1991) notes the differences in meaning between *eye* in the singular and plural.

While it is difficult to predict any such behaviour (or even identify those cases by intuitive means), being aware of any bias would be useful for the overall description of a word. In some cases it is the difference to other, similar, words which is more important than the actual values themselves, as the latter are often influenced by the type of corpus under investigation. A corpus of novels will probably have more verb

forms in the past tense than a corpus of spoken conversations. In written language in general past tense forms are likely to dominate, as they would be the most natural tense for reporting events. However, that prediction is based on assumptions which might turn out to be wrong.

Biber *et al.* (1999, 459) state that *[m]any verbs have a strong association with either present or past tense*, and give some examples of verbs that are predominantly used in only one of the two. However, they separate tense from aspect and only look at simple past *vs* simple present.

Applied to a list of words the tense recogniser will return a string which encodes the grammatical features of the phrase. The possible return values are given in table 4.3.

NounSingular	Pres	Past	Fut
NounPlural	PresPass	PastPass	FutPass
Inf	PresCont	PastCont	FutCont
InfPass	PresContPass	PastContPass	FutContPass
NonFinite	PresPerf	PastPerf	FutPerf
	PresPerfPass	PastPerfPass	FutPerfPass
	PresPerfCont	PastPerfCont	FutPerfCont
	PresPerfContPass	PastPerfContPass	FutPerfContPass

Table 4.3: Possible tense/aspect/number/voice combinations

Two of these values denote noun phrases, and are used to distinguish between nominal and verbal uses of words which can both be nouns and verbs. During the processing stage the individual component counts can easily be extracted from the feature names, and feature combinations are also directly available.

Most of these feature combinations are inextricably linked to the actual word forms, for example only forms ending in *-ing* can be in the continuous, whereas the base form is

much more flexible. Therefore this analysis needs to be carried out on the concordance lines for all inflected variants of a word.

There are some problems: the analysis relies on the output of the word class tagger, which pre-processes the concordance lines before they are fed into the tense/aspect recogniser. The rate of correctly assigned word class labels of modern taggers is about 97%, and with words which are verb-noun ambiguous it can sometimes be difficult to find the correct tag. Especially with bare forms, i.e. non-third person singular present tense, there is not always enough contextual information available to decide on the right analysis without a more or less complete syntactic parse.

However, the errors can go both ways, and a preliminary check of the word *doubt* shows that some nominal uses are counted as verbal ones, and vice versa. As we are only interested in the quantities, rather than the actual instances, it is likely that inaccuracies in the processing will not have a significant impact on the overall outcome. Other inflected forms are far less error-prone, as they are either less ambiguous in the first place, or have added elements in the verb groups, for example auxiliaries or modals.

As an example, for the lemma DECIDE we get the following result with the corpus of 19th century novels (table 4.4):

875	Past
289	Inf
264	Pres
119	PastPerf
96	PresPerf
65	non-finite
42	UNKNOWN
32	Fut
19	InfPass
10	PastPerfPass
3	PresPerfPass
3	PastCont
1	PresCont

Table 4.4: Tense distribution of DECLINE in the corpus of 19C novels

Predictably, the past tense is by far the most frequent tense in a corpus consisting of novels. For the written part of the BNC we get the following result:

7820	Past
3859	Pres
2512	Inf
1614	non-finite
937	PastPerf
840	PresPerf
255	Fut
234	UNKNOWN
159	InfPass
89	PresPerfPass
62	PastPerfPass
56	PresCont
28	PastCont
5	FutCont
1	PastPass
1	FutPerfPass

Table 4.5: Tense distribution of DECLINE in the written part of the BNC

Here we also have the past tense as the predominant tense, but not to the same degree as in the novel corpus. If we contrast this with a third corpus, the Cobuild BBC corpus, we find the same picture again, but this time the more complex present perfect has pushed the present tense out of the top three:

1766	Past
1009	PresPerf
937	Inf
713	Pres
238	PastPerf
213	non-finite
126	Fut
97	UNKNOWN
89	InfPass
55	PresPerfPass
29	PastPerfPass
28	PresCont
2	FutCont
1	PastCont

Table 4.6: Tense distribution of DECLINE in the BBC corpus

If we now compare this with another word, the lemma MEET, from the same corpus (BBC), we get a slightly different picture, reflecting the different ways that MEET is used. The 'UNKNOWN' element here represents mostly the nominal uses of *meeting*. We can see that the past tense is still the most common, but it is closely followed by the infinitive.

8306	UNKNOWN	63	non-finite
2996	Past	62	PastPerfPass
2865	Inf	61	PresPass
2005	Pres	42	PastCont
641	Fut	23	FutPass
539	PresPerf	18	PresContPass
150	PastPerf	8	FutPerf
138	PastPass	5	PresPerfCont
111	PresPerfPass	2	PastContPass
80	PresCont	2	FutCont
76	InfPass	1	FutPerfCont

Table 4.7: Tense distribution of MEET in the BBC corpus

Of course, those numbers give us a sense of the usage of the verbs in question, but it cannot easily be seen how that can be used to gain information of a more general kind. In order to determine the value of the statistic we briefly investigated the distribution of three lemmas across a number of corpora using correspondence analysis (see

section 3.2.4.2 on page 99 for a description of correspondence analysis). The lemmas (randomly) chosen were DECIDE, WALK and BRING, and the distribution was investigated across the BBC, the written part of the BNC, FLOB, FROWN, and the 19C novels corpus.

The correspondence analysis correlates the features according to the profile in the data table, thus compensating for any differences in actual frequency. The features and data items are then projected on to a two-dimensional map, which allows a straightforward analysis of similarities between both features and items, and of the criteria chosen for the similarity of items. The output of the correspondence analysis is shown in figure 4.3.

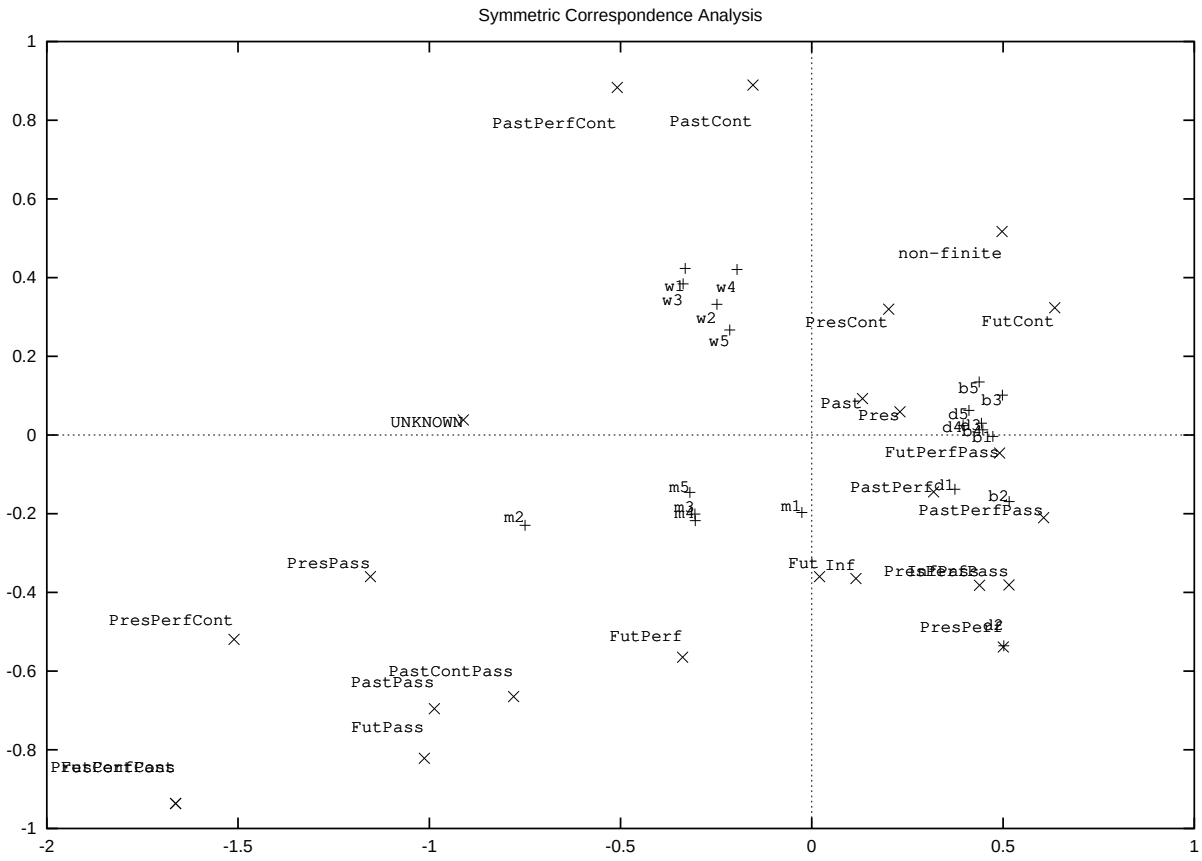

Figure 4.3: Tense/aspect/voice sample

The words are represented by their first letter, **d**ecide, **w**alk, **b**ring, and **m**eet; the distribution of all inflected forms has been used. The different corpora are numbered as follows:

1 - 19C

2 - BBC

3 - BNC

4 - FLOB

Here we can see that for the small sample the tense/aspect distribution seems to be specific to the word, rather than the corpus. The five data points (one for each corpus) of all four lemmas cluster together quite tightly, apart from the ones of the BBC corpus. These represent outliers, though they share the general tendency, in other words despite being further removed from the other four points they are still reasonably close to their respective clusters. This could indicate that there is a mode influence, where the distribution is different in written or spoken language. The BBC corpus contains mostly written-to-be-spoken data, and is thus distinct from the other purely written texts.

The four lemmas are more (in the case of WALK) or less (in the case of MEET) together; DECIDE and BRING seem to have more in common with each other, as their data points almost form a single group. MEET seems to be used differently in 19C novels and the BBC corpus, as those two data points are slightly further away from the others. It appears that in the 19C novels MEET is more frequently used in the passive voice or the present perfect continuous than in the other data sets. The two lemmas MEET and WALK are pulled towards the 'UNKNOWN' element, which reflects their nominal uses.

Interpreting the graph, we can see that the x-axis distinguishes between present perfect, future, and infinitive (on the right hand end) and mainly passive and continuous forms on the left hand side. The y-axis again separates present perfect (bottom end of the graph) from future perfect and past perfect (top end). Broadly speaking, the tendency for the BBC corpus is to use more present perfect than the other corpora, while DECIDE in general is mainly used (in comparison to WALK and MEET) in the present perfect, future, or infinitive. For a more detailed analysis it might be worth separating out the individual features of tense, voice, and aspect to investigate a larger

sample of both lemmas and corpora.

In conclusion we can see that even from only a few examples we can get a useful overall idea of how the features of tense, aspect, and voice are distributed across different word forms and corpora.

4.1.5 Lexical Statistics: Summary

There are two main purposes in computing lexical statistics: the first is to get a general overview of a word's behaviour. While some properties in themselves provide useful information, others become meaningful only when compared to those of other words, so the second purpose is to get an idea of how properties are distributed across different words. In current corpus research it is mainly frequency of words that has been calculated, and frequency is clearly of central importance in analysis.

Many features have not really been investigated. Halliday and James (1993) have demonstrated that the feature distributions of certain (binary) systems in functional grammar divide into either 50:50 or 90:10; this they relate to particular entropy values. Knowing more about the distribution of other features will clearly be beneficial to our knowledge and understanding of language.

4.2 Collocation

Collocation is one of the most established concepts in corpus linguistics; after concordance lines it is the most widely used tool for the analysis of lexis. This is because a word has an obvious impact on the words which tend to co-occur with it, as there is

a close link between context and meaning. Here it does not matter whether the word influences its environment or whether the environment influences the word, as long as we accept that words are not distributed randomly in a text.

For that reason, collocates have been mainly used to approximate the description of the meaning of a word form, and Stubbs (2001) gives several examples of this. In the current project, collocates are also computed routinely as a starting point for further analysis.

In this section we will first try to define collocation as a mechanical procedure; surprisingly there is no standardised way of computing collocations, even though they are so widely used. It is not even common for researchers to mention which procedure they have used themselves when calculating collocations. After the definition we will set out the basic procedure adopted here, followed by a discussion of problematic issues and open questions.

4.2.1 Definition

We can trace collocation back to at least Firth (1957, 11) and his famous remark that *[y]ou shall know a word by the company it keeps.* A large body of literature has developed as a consequence of the central role of collocational analysis in the contextual school. Initial work (e.g. Berry-Rogghe 1973) focused on how to apply traditional statistical methods (such as z-score and χ^2) to corpus linguistics, or was trying to establish independent methods (e.g. Sinclair and Jones 1974), sometimes with a very basic approach to mathematics.

Unfortunately researchers now use many different definitions of *collocation*. They

range from basic (adjacent) two word combinations, or *word bigrams* , to general multi-word units (or *n*-grams), to what we could call the most generally used definition within corpus linguistics, the co-occurrence of two lexical items within a certain stretch of text (the so-called *span* or *window*).

Firth (1957, 12) defines collocation as follows:

> Collocations of a given word are statements of the habitual or customary places of that word in collocational order but not in any other contextual order and emphatically not in any grammatical order. The collocation of a word or a 'piece' is not to be regarded as mere juxta-position, it is an order of *mutual expectancy*.

Monaghan (1979, 32) paraphrases this as *a syntagmatic relation of lexical items to each other in terms of their likelihood of cooccurrence*, and gives Firth's examples of *young ass* and *old fool*, where one could interchange the adjectives, but normally would not do so. The relationship between these two pairs of lexical items thus overrides the general freedom in the relationship adjective-noun. However, this requires further qualification, as one can still say *young fool* in any event: it is a matter of preference rather than a fixed and absolute law.

Such pairings need need to be related to observable context: one can always invent a context in which almost any utterance no matter how weird and unusual could make sense. This point constitutes a vital argument against the use of introspection for linguistic description, since one can prove anything by reference to an individual speaker's judgement on isolated sentences; and thus it means nothing.

The concept of collocation allows us to state lexical relationships in terms of sets of words which have a certain likelihood of occurring near each other. While we would prefer to have fixed rules to describe possible and impossible combinations, Monaghan (1979) argues that one cannot easily formalise this aspect of language, not even by using semantic features or more general categories such as 'animate' or 'abstract'.

Kjellmer (1987, 133) defines collocation as *a sequence of words that occurs more than once in identical form ... and which is grammatically well-structured.* This is (unnecessarily) limited in that it requires grammaticality, but also more flexible as he talks of a sequence, rather than a simple pair of words. However, admitting more than two elements brings us into the area of phraseology, which is better modelled using other constructs (like units of meaning, frames, and chains, as described below). Collocation is used by the majority of researchers as a binary relationship between words, and the most varied aspect seems to be the distance between the elements. It is unfortunate that Kjellmer chose to use the term 'collocation' for what is more appropriately termed 'phrase' or 'group'.

Along similar lines, Graddol *et al.* (1987) list collocation as a syntagmatic relationship under 'sense relations'. They give the example of *white*, which changes its meaning according to the word that follows it, e.g. *coffee*, *skin*, or *wine*. In each case it refers to a different colour. They also include under collocation other syntactic relations, such as the one between subject and verb. We will address this in section 5.3 on 'usage patterns'.

In the (late) 1980s research in collocation made its way into large-scale language processing as part of the then current statistical approaches to language in applied areas such as speech recognition. Collocations were suggested by Church and Hanks (1990) as a possible way to disambiguate between different candidate words proposed by a

speech recogniser or a spell-checker, e.g. deciding whether *farm* or *form* would be the appropriate spelling in a certain context. Around the same time collocations also made their way into lexicography, through the Cobuild project (Sinclair, 1987) and also work at AT&T (Church *et al.*, 1991).

In the field of computational linguistics collocation was unknown until the publication of Church and Hanks (1989), which Manning and Schütze (2000, 187) refer to as *[o]ne of the first publications on the discovery of collocations*, a further example of something being independently discovered in separate fields at different times because of a lack of interdisciplinary communication. The publication of Church and Hanks' work sparked off a large number of further experiments and investigations into collocations and their applications. The lack of progress with purely structural/formal approaches got researchers interested in more statistical and lexicalised methods, facilitated by the increasing availability of large corpora in the 1990s.

Some approaches listed by Manning and Schütze (2000) would not really be accepted as collocational by most corpus linguists: for example, Justeson and Katz (1995) use a number of part-of-speech templates to collect bi- and tri-grams to extract technical terminology. These templates (e.g. 'adj noun', or 'noun prep noun') restrict the results to grammatically useful ones, but otherwise their approach is more like the 'chains' discussed in section 4.3.2, or the work of Kjellmer described above.

A lot of work in computational linguistics is concerned with the exploration of different significance measures. Church and Hanks (1989) propose the use of *mutual information*, a measure derived from information theory, and *t-score*, which is based conceptually on the t-test for equality of different sample means. Since then, a number of other measures have been introduced, though it is not always clear why one approach

is superior to another. A general problem is that some words co-occur frequently with a given word because of their own high frequency; most of them are function words. Furthermore, words in texts are not distributed randomly, although virtually all significance measures assume they are.

Ideally researchers are interested only in content words, so many newer measures attempt to weed out the words with higher frequencies. But the problem then is that there is no clear-cut boundary between function and content words in terms of frequency; instead there is only a tendency for content words to be less frequent than function words.

Collocation is used for a variety of purposes, such as grammatical disambiguation and machine translation. Smadja's Xtract (Smadja, 1993) is a program designed to extract collocations from texts. However, his definition of collocation also diverges from the one traditionally used in corpus linguistics, and is focused more on grammatically linked words, such as 'predicative relations', 'rigid noun phrases', and 'phrasal templates'.

In corpus linguistics, collocation is predominantly used for lexical description, for example in lexicography. There is also work concerned with general properties of collocation as a method of analysis, e.g. Stubbs (2001). Stubbs uses the Cobuild Collocations CD ROM Cobuild (1995) as a database for his analyses. This is a large data source, but it is static in the sense that it has been created for one particular corpus, and its usefulness for comparative studies is thus limited. It is also unclear how the collocations were arrived at, as documentation on technical details is not supplied.

For the purpose of this project we define collocations as *pairs of words where one*

word (the 'collocate') occurs in the environment of the other word (the 'node word') in a significant way. Thus collocation is a purely lexical relationship between two words and is not influenced by any syntactic issues such as well-formedness, and there is also no *a priori* limit on the distance between the two words. This definition, however, involves defining several further concepts in order to be able to implement collocation as a procedure:

word: a word is defined through the use of a tokeniser in the context of this project. Possible modifications are case-folding, lemmatisation, and word-class disambiguation. These will be discussed below.

environment: the environment, or *span*, of the word limits the extent to which it influences the choice of words around it. This can be determined through an automated procedure described below.

significant: significance is a contentious issue with regards to collocation. Since collocation is predominantly an exploratory procedure (though Church and Hanks (1990) try to link it to word association experiments) we cannot say in advance what makes a collocate significant or not. So far this problem has been addressed mainly through an (arbitrary) metric to compute a significance score for a given word pair.

We can now summarise the basic procedure of computing collocations as follows:

1. locate all instances of the node n in the corpus.
2. construct a subcorpus of the words in the environment e of n.

3. create a frequency list *f-e* of the words within *e*.

4. compare *f-e* with a representative frequency list *f-r*.

5. assign to each word in *f-e* a significance score based on the comparison.

6. sort *f-e* in descending order of significance.

This description does not suffice as an implementable algorithm, as too many concepts remain underspecified. In the following section we will discuss those concepts in more detail and will suggest a 'prototypical' specification of the collocation algorithm.

4.2.2 Parameters

Most published work on collocations does not state what parameters are used for calculation, and often such information is not even accessible in the software used for the process. That means that such studies cannot be replicated, even if other elements (e.g. the corpus used for the investigation) are documented and available. So it is very important to be explicit about every aspect of the algorithm, and in this section we will summarise what the relevant parameters are.

The above definition of collocation as it stands is basic and imprecise, because several concepts still need to be specified further. Mason (2000b) provides a list of parameters involved in the collocational process that need further investigation:

- The choice of corpus: this counts as one of the most fundamental parameters, as lexical variation between different types of corpus data will heavily influence

collocation. Word meanings change with time and place, and different collocations will reflect those differences when using different corpora.

- The choice of the node word: apart from the word itself we could include other variants in the analysis, such as inflected forms, spelling variants, upper/lower/mixed case versions, even semantic classes (e.g. COLOUR; that would be a move towards *semantic preferences*).

- The choice of collocates: as for the node above, we can apply a number of pre-processing steps to the collocates as well, merging or splitting word form frequency counts.

- The window size and shape: how many words either side does it extend, symmetrical/asymmetrical shape, different weightings (e.g. Hann(ing) window, as used in digital signal processing).

- The choice of significance function to evaluate the status/value/worth of the collocate.

- Cut-off value: programs usually ignore collocates below a certain threshold in order to deal with typographical or spelling mistakes, or 'weird' word types.

4.2.2.1 Node/Collocate

In actual corpus data one can find a lot of variation in the spelling of a word, such as upper case characters either in sentence-initial position or for emphasis. We also face the problem of proper nouns which introduce an ambiguity, e.g. *Brown* as either a name or a colour term. Sampson (1995) talks at great length about the manual handling of

such cases; however, we cannot treat ambiguous cases manually in large-scale corpus work of the kind described here.

Dealing with mixed-case texts in fully automatic processing poses a dilemma: we introduce a degree of error whatever we do. Either we treat some names like other words, thus interfering with the distributional statistics, or we treat some 'ordinary' words as two different word types simply because we find an instance in sentence-initial position. Either way, it could cause a serious problem as we have here a systematic variation (rather than a random one), and we cannot assume that a larger amount of data will reduce the size of the introduced error.

The same applies to lemmatisation: previous studies (e.g. Stubbs (2001) on *seek*, and Sinclair (1991) on *decline*) have shown that different inflected forms do not always behave the same. We would expect different behaviour once we take into account that they occur in different syntactic environments (due to different tense or number inflection). Just as with case-merging, we can use lemmatisation to boost the number of occurrences (and reduce the number of word types to process), but we will lose some differentiation between the various inflected forms in the process, and might end up with less precise results. But we might doubt whether we can achieve higher precision at all.

The opposite holds for part-of-speech tagging: whereas lemmatisation and case-folding merge different typographic/orthographic forms to one canonical representation, tagging differentiates between the various forms which have separate word classes, thus allowing a more fine grained analysis. However, if word classes are defined based on distributional features we will simply observe what we have already put in, namely that the noun *light* has different collocates from the verb *light* simply by virtue of being

a noun, which occurs more frequently with adjectives and less with personal pronouns or adverbs.

With collocates a further preprocessing step applies which can under certain circumstances (see the discussion on significance functions below) heavily influence the outcome: the threshold. In statistics one usually requires a minimum number of occurrences of an event in order to accept the statistics as valid, e.g. a χ^2-Square test demands a minimum of 5 instances for each cell in a contingency table. For the processing of collocations no necessary threshold exists, but it would make sense to exclude rare items on the grounds that they are not very relevant because of their low frequency. Furthermore, spelling and other typographical mistakes can easily introduce 'new' wordforms, which might be interpreted as highly significant, as they only occur a few times, and in the same environment.

One problem with a threshold is that it is at least partly dependent on corpus size: a large corpus has many more words, and many more rare words, and also more potential for spelling mistakes being repeated (and thus pushed above any fixed threshold value).

There cannot be an objective value that is determined in advance of calculation, but we should keep in mind that infrequent words might have to be discarded later on as artefacts of technical issues (typesetting, scanning, etc).

4.2.2.2 Environment

We define the environment of the word through a 'window' on the concordance line, the size of which we call *span*. The span of a window describes how far away from the node word it extends, and we present it as two numbers separated by a colon, i.e. `4:4`

for a span of four words to the left and four words to the right. No particular reason exists for a symmetrical span, so `3:5` defines a span of three words preceding the node, and five words following it.

Few researchers mention what span value they use, which indicates that they do not consider it an important parameter. 'Traditional' values, based on Sinclair and Jones (1974) and later Stubbs (1995) remain in use, and either a value of `3:3`, `4:4` or `5:5` is used. Stubbs (1996) mentions that this is a problem which at present has no solution; and irregularities within the orthographic system make it difficult to find one.

The problem is the non-standardised spelling of words like *teapot*, *tea-pot* or *tea pot*, or even *all right* and *alright*. There is no clear match between words and tokens, and it is doubtful that one can ever be achieved. In same cases even a standardised spelling can be tokenised in different ways, for example *don't*: either *do* and *n't* (BNC) or *don* and *t* (Bank of English).

Remarkably in this context, the corpus access software used on the Bank of English does not even document what value it uses as span size for computing collocations. Instead the actual values are hardcoded in the software inaccessible to users. Little actual research on the topic exists, in marked contrast to the choice of a suitable significance function (see below for a discussion of that parameter).

Sinclair and Jones (1974) describe some initial basic research trying to find the optimal value of the span by investigating the number of different word types at different distances from the node. However, they did not have very much data to work on, and also assumed that the span would be symmetric, an assumption disproved in Mason (2000b), where their study was reworked with more data and fewer preconceptions.

One of the results of Mason (2000b) shows that each word has its own ideal span, now defined as the area of influence on the context. We can determine this influence empirically, and it yields a property called *lexical gravity*.

Another assumption states that the span has always got a rectangular shape, i.e. that no difference in weighting exists between the words immediately adjacent to the node word and those at the outer boundaries of the span on either side. In digital signal processing other window types are in common use (mainly to avoid computational artefacts because of pretending that a non-stationary signal has quasi-periodical properties). For the purpose of this thesis, we investigate three different window types:

1. Rectangular: the 'traditional' window with even weighting throughout
2. Triangular: the weight decreases linearly with distance from the node
3. Hann(ing)[1]: the weight value involves computation using a cosine, which gives a more 'rounded' decrease as compared to the triangular window

In signal processing one typically uses window sizes of 128, 256, or 512 data items, so the difference between triangular and Hann(ing) windows will probably not result in any dramatic differences when dealing with sizes of around 10.

4.2.2.3 Significance

The simplest way of identifying words which frequently co-occur with the node word uses frequency, i.e. one prepares a list of words within the span sorted by their frequency

[1]This window is named after Julius von Hann, but usually named in the literature in analogy to the Hamming window (after Richard Hamming)

of occurrence. However, this list will typically mirror a general frequency list, in that words with an overall high frequency in the corpus will also appear in the upper ranks without having any special relationship with the node word. Almost any noun will have *the* as one of the top collocates for purely syntactic reasons.

The standard solution to this dilemma involves transforming the frequency list with a mathematical function, which calculates a *significance score* for each collocate based on input parameters such as the collocate's frequency and the number of times node and collocate occur together (and separately). This often involves comparing two numbers: the observed frequency derived from the concordances, and the expected frequency, derived from the collocate's overall frequency in the corpus or a reference frequency list.

Berry-Rogghe (1973) reports experiments with collocations using the *z-score*; as another straightforward score one can use the simple ratio of observed over expected frequency. Church and Hanks (1989) introduced *mutual information* (mi), a concept imported from information theory, and *t-score*, a variation on the t-test from statistics. Barnbrook (1996) and Stubbs (1995) list the exact formulae of these scores. Later additions to the list include a *weighted mi-score*, where the reference frequency of the collocate counts more (either squared or cubed, Oakes 1998), and the *log-likelihood coefficient* (Dunning, 1993).

Most significance scores have weak points as well as strong ones, and it requires a lot of practical work with various scores in order to find out how to best use them. Mutual information tends to favour rare words, which frequently include typographical or spelling mistakes. Introducing a threshold or cut-off point will usually solve that problem, but then words whose frequency lie just above the cut-off will tend to score

highest; the threshold value thus has an important influence on the result. The t-score behaves in a less extreme fashion and lower frequency words score less high, but it makes it harder for medium frequency words (which will be important collocates) to break through into the ranks of the words with generally high frequency.

Sinclair (1991) describes an alternative approach to the mathematical transformation: upward and downward collocates, where the difference between a collocate's reference frequency and the node's frequency decides its significance. Significant collocates have a lower frequency value (with a grey area of words with roughly equivalent frequency).

The problem with this straightforward definition of 'upwards' and 'downwards' concerns words which have similar frequencies. It is left to chance whether one word is marginally more frequent than the node word (and thus is ignored as being an upwards collocate) or whether it is slightly less frequent, leading to its inclusion as a downward collocate. A solution to this dilemma is the adoption of frequency bands which allow for a certain margin of variability. Quasthoff (1998) proposes a way of computing frequency bands (see section 4.1.1). Frequency bands add an element of vagueness, as two words whose frequencies are very similar would end up in the same frequency band, despite small differences. This allows for variability due to chance.

This last method of using upwards and downwards collocates has one important methodological advantage over the transformation-based significance scores: it makes no claims of significance based on mathematical grounds. We still know so little about the distributional properties of lexical items that it is hard to justify the technical use of 'significant'. In fact, Baayen (1991) classes linguistic events as 'large number of rare events' (LNRE), which are difficult to describe with commonly used mathematical mod-

els. About 50% of words in a corpus will occur only once, and similar distributional observations have been made for production rules in a phrase structure grammar (Sampson, 1987). So we can say with confidence that linguistic events do not follow a normal Gaussian distribution, and we cannot apply any statistical procedures which require this distribution, a caveat that many researchers easily forget when confronted with 'significant' results for virtually anything they look at.

4.2.2.4 Threshold

As mentioned previously, a threshold or cut-off point is often introduced to deal with words that are deemed to be insignificant due to their low overall frequency. They could be genuinely rare words for which not enough data is available for statistically 'significant' results, or errors introduced by spelling, typing, or data transfer (e.g. optical character recognition).

There is no scientific reason for choosing any particular value, and one can expect that by excluding tokens with frequency n the focus then falls on tokens with frequency $n+1$: here too there will be mistakes, only they were not noticeable before because of the mistakes now discarded. The higher the threshold, the higher the likelihood that 'legitimate' words will be ignored, so it is difficult to make the right decision. This is especially important when dealing with significance functions which boost very low frequency words (since they have the highest information content). Other scoring functions are more robust and are not so easily affected.

The upward-downward method is 'immune' to this problem, as it only uses raw frequency (or frequency bands). 'Weird' tokens will still be at the bottom of the list, and if processing starts from the top all the 'proper' tokens will receive attention first

before those that are further down the list where the distinction between mistakes and rarities becomes difficult. A threshold thus becomes relevant only when other means of evaluating the outcome are used, and when significance scores result in infrequent words receiving a boost to score higher than more frequent ones.

We have here hit upon a general dilemma in corpus linguistics: what to do with rare events. Since Baayen (1991) found that language consists of a large number of rare events, ignoring infrequent elements quickly means that we narrow down our scope too much. If half the words in a corpus occur only once, then a threshold value of two will reduce coverage by 50%. On the other hand, we can hardly make sensible statements on the basis of single occurrences.

Realistically, despite accounting for a large number of occurrences overall, rare events are not that important for the description of language. After all, they do occur only rarely, and are thus of limited value to a language user.

Therefore the choice of threshold value remains arbitrary. It could either be fixed (e.g. at 5), or the value could depend on corpus size. For the latter a formula such as the following might be appropriate:

$$THRESHOLD = \sqrt{\frac{N}{1,000,000}} \tag{1}$$

where *N* is the size of the corpus in tokens. For a corpus of one million words that would put the threshold value at 1, for ten million words at 3, and for one hundred million words at 10.

However, these values are still arbitrary, and ultimately the appropriate settings depend on the particular application.

4.2.3 Lexical Gravity

After a brief discussion of relevant parameters, we will now investigate the possibility of determining empirical values for at least the window size. Previous work has shown that it is possible to define objective criteria to identify an ideal value.

Mason (1997) introduces the concept of *lexical gravity*, which is further refined in Mason (2000c). It has been developed in an attempt to identify an objective criterion for determining the ideal span setting for processing a word's environment, such as computing collocations. Span is one of the many parameters involved in the collocational procedure, and it has a potentially large influence on the outcome.

In one extreme case, the span would simply be 1:0 or 0:1, i.e. just the preceding or following word. The outcome would be mainly determined by the word class: a noun would have determiners, adjectives, or verbs, but other nouns only via nominal compounds. The collocate would be part of the same phrase, or of the immediately adjacent phrase in the case of boundaries (such as a span of 1:0 with a determiner as the node). The span 0:0 is also possible, in which case a word would not have any restricting influence on its environment, and therefore no collocates at all.

There is no upper limit for the span, though it quickly becomes pointless once the value is too large: at a certain point the influence the node word has on its environment is overshadowed by the influences of other words. This is not to be confused with the notion of long-distance dependency drawn from theoretical linguistics: while a word's

actual influence will affect the element even beyond a measurable distance, it would only influence certain grammatical features, such as number or gender for agreement purposes. When looking at collocation, that phenomenon can be disregarded.

Lexical gravity can be used to determine a setting for the span width. If we assume that the span should extend as far as the influence of a word on its environment, then the width of the span corresponds to the area around the node word where the lexical gravity is different from its standard value. We can thus specify an objective way to determine a span value which is empirically motivated.

The procedure involves a notion which is hard to define in computational terms, namely the 'standard value' of the lexical variability. In Mason (1997) it has been shown that the variability in a text is very uniform, i.e. that there are no points (within the first 500 words of a text) where the position of a token within the text has an impact on the possible lexical choices. When looking at a number of different texts we are as likely to get as many different words at position 50 as at position 123. But how can we identify this value in a set of concordance lines?

The first approach would be to take the average variability value (represented here by the negative entropy). However, the concordances differ from a set of texts in that they have a lack of variation in the node word position, which will lower the average computed; even excluding the node position from the calculation distorts the result, as the positions that are eventually to be placed within the span have lower variability values. A solution is to use the median instead of the average, which is more stable.

We then need to implement the 'different from' property. The median value itself is not suitable, as there is always a certain degree of variation in the actual values, so

we need to set as the target value a range around the median, for which 0.2 has been chosen. With a reasonably large data set the entropy values tend to be around 9.2 or 9.3, so that we need to allow for some fluctuation.

The procedure starts with a span of 0:0 and moves outwards to both the left and the right hand side. If the entropy value at the position in question is within 0.2 of the median the procedure is stopped, otherwise the span value is incremented and the procedure continues. Should the variation be larger than the median the procedure also continues, as observations (Mason, 1997) have shown that certain types of words (e.g. determiners) can have a 'negative' gravity.

The resulting span values are stored in a file for later use by the collocation procedure.

Looking at the distribution of span sizes as calculated through the lexical gravity procedure we can see the following, rather surprising, trend: the distribution of span size roughly follows a bell curve, with the mean at 4. As we cannot have a negative span width (in fact, it also has to be non-null) the left-hand side of the curve is truncated at the width of 1, while the right-hand-side expands as far as 16, which occurs just once. All word forms with a frequency of less than 10 were excluded on the grounds of insufficient data; from about 25,000 word forms from the Cobuild head word list that occurred with a frequency of at least 3, there remained 17,500 forms. Of these, 3,346 forms have a span width of 4. The distribution is shown in figure 4.4.

The reason why this result is surprising is that hardly any linguistic data has a normal distribution; power-law distributions (such as Zipf's law) are much more common.

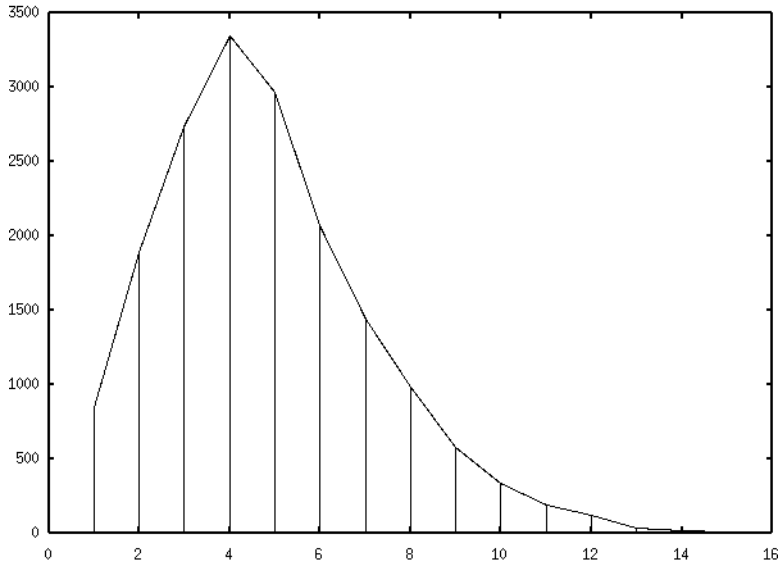

Figure 4.4: Span distribution in the BBC corpus

There is no correlation between word frequency and span width.

The next graph (figure 4.5) shows the distribution of the individual span values. The left span is represented by negative values, whereas the right span is positive. Neither the left nor the right span can cross zero, as the node word is always included in the window. Each word thus has two data points in the graph, one for the left boundary, and one for the right.

We can see that the most frequent span values are 2 to the left and also 2 to the right, and the overall distribution appears to consists of two merged bell curves. Again, due to the node word boundary the two curves are truncated on one side. We cannot say from this graph that the most frequent span will be 2:2 (though it looks likely), but the symmetry is remarkable.

If we look at a frequency list of the top span values (table 4.8), we notice that 2:2 is indeed the most frequent. With about 500 fewer occurrences 2:1 and 1:1 are next.

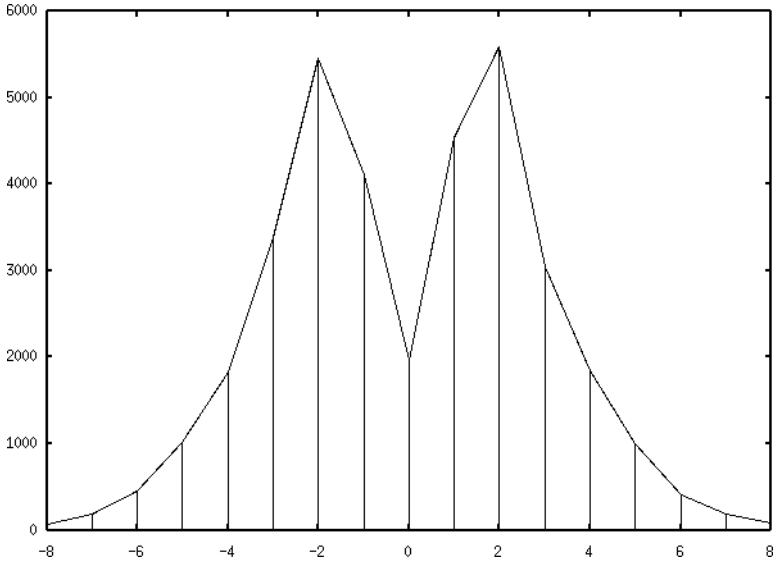

Figure 4.5: Frequency distribution of span boundaries in the BBC corpus

1899	2:2	633	3:3
1390	2:1	603	1:3
1361	1:1	585	4:2
1163	3:2	513	2:4
1147	1:2	451	0:1
1032	2:3	400	3:4
749	3:1		

Table 4.8: The frequency of the most common span values

Obviously, the span width of 4 achieves its high frequency through the additional combinations that result in a size 4 span, such as 3:1. The frequency distribution of span values is shown in figure 4.6. The horizontal axis here represents the individual span values as shown in the table above (and sorted by descending frequency), whereas the vertical axis shows their frequency.

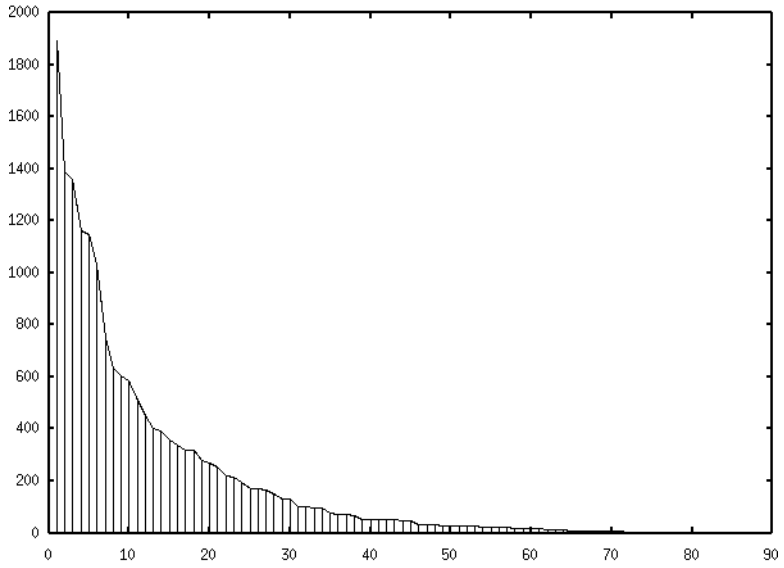

Figure 4.6: Frequency distribution of individual span values in the BBC corpus

4.2.4 Collocation Post-Processing

A straight list of collocations is of limited value as a description of a word form. In order to extract more information from the procedure, several post-processing steps need to be taken which can increase the descriptive power of collocations.

Stubbs (2001, 27) provides a sample analysis of SEEK, illustrating that its different inflected forms do not share all the same collocations. He states that the *overlap in their collocates gives us one measure of the semantic distance between the word-forms* (2001, 28). Calculating this overlap is therefore the first step in post-processing.

We next need to identify the 'lemma-collocates', which are those collocates that are shared between all inflected variants. In a way these lemma-collocates reflect the core meaning of the lemma. If there are none, then the lemma just happens to be an umbrella for several distinct forms which are not really related semantically, though they may

have been in the past. It is important to keep in mind that language is a fluctuating system, where relationships between elements shift at different speeds and are therefore at different stages in their development. We cannot assume that a formal relationship between some classes of elements will hold for all relevant items.

We can then determine the distances between each form and all other forms: this will be the ratio of the number of shared collocates compared to the total number of collocates the two elements have. This distance is not necessarily symmetric, as $form_1$ could share all its collocates with $form_2$, whereas $form_2$ has a few additional non-shared collocates. Thus $form_1$ will contain a subset of the joint collocate set, whereas $form_2$ contains the complete set. Then the distance from $form_2$ to $form_1$ will be 0.0, whereas the distance from $form_1$ to $form_2$ will be, say, 0.2. Assuming C_1 and C_2 for the sets of collocates of $form_1$ and $form_2$ respectively, we get the following relationships:

$$d(f_1, f_2) = 1 - \frac{|C_1|}{|C_1 \cup C_2|} \tag{2}$$

$$d(f_2, f_1) = 1 - \frac{|C_2|}{|C_1 \cup C_2|} \tag{3}$$

An alternative measure could simply take the amount of overlap, in which case it would be symmetrical. This is advantageous from a mathematical point of view, as we can then assume Euclidean geometry of the semantic space, but it is questionable whether the result will then be adequate in describing observations from psycholinguistics (see chapter 6 for a discussion of issues relating to semantic space). An alternative measurement would be:

$$d(f_1, f_2) = 1 - \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|} \tag{4}$$

All the measures are ranged between 0.0 and 1.0, with a smaller number representing a closer similarity. If C_1 and C_2 are identical, both terms will be the same and thus d will become 0.0. If there is no overlap at all, the fraction will become zero, and d will end up as 1.0.

There are a multitude of similarity measurements available, developed for a variety of purposes. Lin (1998) lists a number of measures which are commonly used in information retrieval. For our analysis we can consider the cosine, Dice, and Jacard measures:

$$cosine(f_1, f_2) = \frac{|C_1 \cap C_2|}{\sqrt{|C_1| \times |C_2|}} \tag{5}$$

$$Dice(f_1, f_2) = \frac{2 \times |C_1 \cap C_2|}{|C_1| + |C_2|} \tag{6}$$

$$Jacard(f_1, f_2) = \frac{|C_1 \cap C_2|}{|C_1| + |C_2| - |C_1 \cap C_2|} \tag{7}$$

Using the results of this analysis we can perform a cluster analysis to create a visual representation of word form similarity in the form of a dendrogram. With these results in mind we can explore whether the collocates of a given word are dependent on a particular inflected form, or whether they are valid for all forms of the lemma.

However, while this sounds good in theory, a practical problem is that the inflected forms do not actually have any shared collocates. This means that it is not possible to compute any similarity relations among them, unless a different parameter setting is used for calculating collocations. Using the 'default' settings applied throughout this project, there were no shared collocates for the various form of SEEK, MEET, or HOUSE, using the complete set of collocates found for those lemmas.

When investigating meaning, we will look at using collocations as a means of determining semantic similarity between different words (see section 6.2). In order to achieve some degree of overlap, we will then collect frequencies for the collocates of all candidate words, even if the collocates are not actually collocates of the word close to which they occur.

4.2.4.1 Summarising Information

Throughout his book, Stubbs (2001) uses a convention to compress the display of collocations, where he lists collocates in pointed brackets (<...>), combining inflected forms where this is relevant. An example from page 47 is:

> CAUSE <problem(s) 1806, damage 1519, death(s) 1109, disease 591, concern 598, cancer 572, pain 514, trouble 471>

This list of collocates relates to the lemma CAUSE, and the singular and plural forms of *problem* and *death* are combined, which boosts their respective frequency counts. But this has not been done for *damage*, *disease*, *concern* etc, presumably because the respective plural forms were not frequent enough to warrant inclusion in the list.

A problem thus arises with words where both singular and plural forms are not frequent enough to be counted as collocates on their own, but where combining them would result in them being included. For that reason we need to retrieve the full set of collocates in the initial step before we do the post-processing, in which some lemmas might do better than their individual forms.

The question is when we want to combine inflected variants. On the one hand we might miss a useful generalisation by keeping forms separate; on the other hand we might blur the distinction between forms. Given the example above we could speculate that *damages* are not caused, but rather sued for, whereas both a single *problem* and multiple *problems* can be caused. We thus need a heuristic to decide when to combine forms and when to keep them separate.

Initially we can assume that forms can safely be combined if they have approximately the same frequency of occurrence. That would also be the case where joining them yields the largest increase in rank. If, on the other hand, their frequencies are very uneven, the higher-frequent form does not gain as much as the lower-frequent one, and the risk of blurring the distinction increases. In the absense of any definite guideline we arbitrarily combine forms if the lesser frequent form has more than half the frequency of the higher frequent one. This also applies if there are more than two forms: each individual form to be joined requires half the frequency of the one with the highest frequency.

Alternatively, forms could be joined if a change of rank was involved. In the first step, all forms of a lemma which are included in the set of collocates above the defined threshold are joined; it makes no sense keeping them separate in that case. Then, starting at the bottom of the list of words above the threshold, any inflected forms are added

to candidates if it means that their combined score will be sufficient for them to climb in the ranking of collocates. Later, however, they might drop down again if other words gained through their own inflected forms.

A third alternative would be to include inflected forms as long as they cross the minimum threshold anyway. In the current process there are two thresholds involved: one that filters out erroneous words, *hapax legomena* and spelling mistakes; and a second, much higher threshold that determines which collocates out of the set of significant candidates are presented to the user. The justification for this simplified procedure is that once we assume that multiple inflected forms could be significant collocates, then it could be a mere accident that a particular form is not included in the set. If it does not occur frequently enough even to get past the lower threshold, however, then we can assume that it is not a significant collocate at all.

The merit of the third alternative is that it is simple to implement, and it does not introduce any element of selectivity beyond what already exists. For those reasons we will choose this method for compressing the information given to the user.

If two forms have been joined we can show it by a combined form, such as 'problem(s)' above, or 'seek(ing)'. If there is no clear overlap in orthography they will be given as 'seek/sought'. If all forms of a lemma have been joined we simply give the lemma: 'SEEK'. For more than one form the best solution seems to be 'seek/-s/-ing', which would mean that the forms *seek*, *seeks*, and *seeking* were included. Should the base form be missing we would have to go back to 'seeks/seeking' to make clear the distinction.

This can be done automatically using a simple string matching procedure: the shortest form is taken as the base, the complete set of inflected forms is retrieved from the lemmatiser, and then any characters appended to the shortest form are identified. If no match with the base form is found, then the whole of the inflected form is added.

For the remainder of the description we are left with the absolute frequencies of the collocates co-occurring with the node. Stubbs sometimes expresses them as percentage values, presumably if they comprise a large enough proportion.

An additional issue that cannot easily be resolved in a fully automated way is that of semantically related groups of words, which for example could be labelled as 'colour', or 'vehicle', or whatever group of words happens to occur. This would extend the generalisation from simply morphological (as in the case of inflected variants) to semantical features as well; words to be grouped together would share enough of those features to be treated the same. However, depending on the level of specificity of the description this might not in fact be very useful, as it is easy to generalise too far, and the set of all elements is far less easy to define than in the case of morphological variation.

For the time being the compressed description is thus restricted to morphological variants only, while semantic issues may be explored in future work.

4.2.5 Collocations: Summary

The computation of collocations involves a number of parameters which influence the outcome to a large degree. Unfortunately most studies rarely report the chosen values, which hinders replicability, and thus reduces the overall usefulness of collocation as a scientific tool. Despite a long tradition of working with collocations, which as a concept

was introduced by Firth more than half a century ago, no systematic study of the relevant parameters has yet been undertaken. That is partly due to the vague definition of collocation, and the variety of uses it can be put to. As it is principally an explorative tool, it is not possible to judge in advance what the collocations of a word should look like.

This is a recurrent theme that hinders the evaluation of empirical procedures. We can assess whether the results are plausible, but we cannot decide whether they are correct or not.

For the present work we have tried to justify some choices through further empirical studies, such as the ideal span width through the lexical gravity of a word, whereas others (the type of window, the threshold) are still open to change and preliminary decisions have been made through educated guesswork. Still further choices (significance function) have been made through general theoretical reasoning.

The important point here is to be aware that none of these default values is set in stone. Instead the values used for a particular application should be made transparent and users should be able to adjust them in order to compare results. In the meantime a comprehensive comparative study of the different possible parameters is left as a desideratum for future work. Owing to the number of parameters and the possible value ranges this is beyond the scope of this thesis.

4.3 Multiword Units

When studying language carefully one quickly realises that there exist units that go beyond the traditionally defined word. These units are usually located below phrase level, but behave similarly to individual words, only they are longer. They can be called multi-word units (MWUs), to use a neutral term. By definition a multi-word unit is at least two words long. There is no set maximum length, but in practice recurrent sequences of words are limited: if the unit is too long, it will be too specific to be used in a different context. If the unit occurs only once we cannot be sure that it is really a unit, rather than an accidental usage or an unmotivated string of words.

There are many linguistic phenomena to which the term MWU can be applied; the most commonly known are *phrasal verbs* and *idioms*. MWUs are often defined semantically, i.e. the meaning of the whole unit has to be different from the combined meaning of the single words that make it up. That would exclude straight compounds, such as *satellite dish*, but include phrases such as *a red rag to a bull*. There is also the question of variability; some definitions allow for changes in word order or inflected forms.

On the other hand, the term could be applied to any sequence of words that is used as a building block for an utterance. The latter definition has the advantage of not requiring human judgement, as it does not rely on meaning, and can thus be more easily automatised. While suitability for computer processing alone should not be seen as a defining feature of linguistic units, related work (Dias, 2003) mentions two reasons why MWU identification should be automatic: 1) to be language independent, and 2) to be able to discover different types of multi-word units.

In the end the appropriate definition depends on the purpose. The purpose of the automated identification described here is to discover the phraseology of a word, i.e. the typical ways in which a word is used in context. A multi-word unit is here defined as a recurrent combination of two or more words. However, not all such units are of interest, as many will be simply be caused by arbitrary syntactical restrictions. For example, most nouns will be preceded by *the* many times, but that does not mean that it is an interesting combination. It might, on the other hand, be seen as an intermediate unit between word and phrase.

There is a fundamental problem with the identification of multi-word units: it seems sensible to postulate such units, but little can be said about them in advance regarding their length or form, without introducing an unacceptable bias to the analysis. The process of finding multi-word units is therefore an exploratory one, where we look for something we expect to exist, but without being able to state explicit criteria for success or failure.

4.3.1 Related Work

The identification of multi-word units belongs to a rather ill-defined area of language research. Frequently (e.g. Schone and Jurafsky 2001) those units are called collocations, even though this term is traditionally used (at least within corpus linguistics) only for free co-occurrences of a node word and a single collocate (Sinclair, 1991). The next open question is that of syntactic well-formedness: Dias (2003) links multi-word units to recurrent syntactic patterns. This is similar to Kjellmer (1987) mentioned above, who requires that collocations should be well-formed units.

Merkel and Andersson (2000), referring to Smadja (1993), state that multi-word units are domain-dependent; however, that does not apply to those units made up of very frequent words. Only units which predominantly consist of domain-specific content words would be domain-dependent themselves.

Multi-word units of fixed length are often called *n*-grams. Stubbs (2003) looks at a number of differently sized *n*-grams, calling them *chains*. This term will be used to refer to one type of multi-word units based on variable-length *n*-grams later on.

Renouf and Sinclair (1991) describe what they call *collocational frameworks*, a pattern with a blank slot surrounded by high frequency words, such as *as* ___ *as* or *the* ___ *of*. The number of content words that fits into such patterns is often quite limited. The patterns are pre-determined, i.e. set by the analyst. Arguably the resulting filled frameworks can be viewed as multi-word units.

We will now look at two approaches to the recognition of MWUs: *chains*, based on *n*-grams, and *frames*, which are similar to Renouf and Sinclair's frameworks.

4.3.2 Chains

Chains as defined by Stubbs (2003) are recurrent *n*-grams with differing lengths starting at a minimum length of two words. His study shows that chains of high frequency words occur much more frequently than would be expected by chance, providing insights into the phraseological patterns of English. Stubbs and Barth (2003) use a list of such chains to investigate the distribution of different multi-word units across different texts or corpora; but chains are less useful for describing the phraseology of a single word. Stubbs and Barth do not claim that these units are linguistic units; they only use them

for diagnostic purposes on text types.

In this section we will present an extension of this method which attempts to process the results automatically. Unlike Stubbs we do not select the *n*-grams from the complete corpus in sequential order, but instead from the context of a given word. We thus treat chains as a feature/property of a *word*, rather than a *text*, even though the outcome would be the same if all words of a corpus were processed.

The chains of a node word are the set of *n*-grams containing the word, with *n* starting from 2 (for bigrams) and going up to about 7. For longer *n*-grams the frequency counts are usually too low to yield any more repeated sequences, and we could see 7 as the upper limit with support from psychological research (Miller, 1956). The position of the node word within the *n*-gram is fully variable.

In order to retrieve multi-word units starting from a specific node word, the procedure does not need to work on a full corpus, but rather on a subset centred on the word in question. This subset can easily be collected using concordance lines. The procedure for collecting chains is then straightforward: from a set of concordance lines of the node word we start at position *n* on the left hand side and proceed to position zero (i.e. the node word itself), adding the *n*-gram starting at that position to our list. As we are interested in frequency counts we need to keep track of the number of times each chain occurs. The list of chains is then sorted according to frequency, and we can use the top of the list as a guide to the phraseology of the node word in respect to fixed phrases; for variable phrases this method will not work, as the frequency counts will be too dispersed, but fixed phrases which are frequently repeated will rise to the top of the list.

One basic problem with the resulting list is that it contains a number of *n*-grams of different lengths, and there will be a lot of overlap. For any trigram there will be two bigrams which overlap, and these bigrams will have at least the same frequency as the trigram, and more likely a higher one. For that reason frequency alone is not suitable as a filter, since longer (and more specific and interesting) chains would then be discarded. We need somehow to filter out the short overlapping chains to get to the really interesting ones. But there is a trade-off between length and frequency: longer chains are rarer (but potentially more interesting), while shorter chains are more frequent (but usually less meaningful).

Kjellmer (1984) is faced with a similar problem in his work on collocations (his use of the term would more correctly refer to grammatically well-formed bi- and tri-grams, see section 4.2 above). He is interested in evaluating the 'distinctiveness' of sequences, and suggests the following (mostly binary) criteria: absolute frequency (more/less than 3), observed/expected frequency, length of sequence (two/more than two elements), textual distribution (single text/multiple texts), distribution over text categories (of the Brown corpus), simple/complex structure.

The criteria he suggests are reasonably effective on a small corpus which is extensively labelled for text categories and structural complexity (He uses an annotated version of the Brown corpus). For larger corpora his threshold values would be far too small to allow his method to be used as a filter. Furthermore, his requirement that sequences should be grammatically correct does not apply to arbitrary *n*-grams, which makes the final criterion inapplicable.

Kita *et al.* (1994) have devised a *cost function* (according to Oakes (1998, 188)) which tries to work out an indicator of whether to select the longer or the shorter of

two overlapping phrases. Their work was oriented towards identifying collocations, but the cost function can be applied usefully to the filtering of chains. Given two overlapping sequences *a* and *b*, with *b* being the longer one, we define the cost function as

$$K(a) = (|a| - 1) * (freq(a) - freq(b)) \quad (8)$$

where a small value of $K(a)$ indicates that the shorter sequence *a* should be selected, and a larger value that *b* is to be preferred. Oakes (1998) does not mention any cut-off values.

The interpretation of the cost function is difficult, as its values depend on the frequencies of the two sequences involved. If both values are low there will be less variation, and a low cut-off point will work best. If, however, the respective frequencies are high, there will be much more scope for variation and we would want to tolerate even larger differences as not significant. For that reason we have defined the threshold as a percentage: if $K(a)$ is less than ten percent of the frequency of the shorter sequence we will discard the shorter chain. If it is more than a quarter we will instead discard the longer one. For values in between we do not take any action and keep both chains.

The procedure is implemented using a tree structure where each word is a node, so that *n*-grams sharing the same initial words will be on the same branch of the tree. This is useful for pruning the tree using the cost function. Initially all *n*-grams are added in back-to-front order, and the pruning starts from the end of the *n*-grams. After the first pruning step the tree is reversed (and the *n*-grams thus appear in the right order) and the pruning is repeated. This procedure enables us to discard the sequence *spite of the* in favour of the more desirable *in spite of the*. In some cases we will still not be able to

perform a proper comparison, namely when the longer chain has additional elements at both ends. But these will generally be a lot less frequent than the 'core' chains.

For example, looking at the chains for *spite* in the BBC corpus (frequency 905 occurrences) we get the following list (unfiltered, i.e. without applying the cost function):

899	[spite] of	25	in [spite] of its
690	in [spite]	24	[spite] of his
689	in [spite] of	22	[spite] of their
265	[spite] of the	21	[spite] of an
206	In [spite]	20	[spite] of the fact
206	In [spite] of	20	[spite] of the fact that
197	in [spite] of the	20	in [spite] of an
76	that in [spite]	19	in [spite] of this
76	that in [spite] of	18	In [spite] of this
67	In [spite] of the	17	said that in [spite]
61	[spite] of a	17	said that in [spite] of
55	in [spite] of a	16	[spite] of that
38	But in [spite]	16	in [spite] of all
38	But in [spite] of	16	says that in [spite]
37	[spite] of this	16	says that in [spite] of
32	[spite] of its	15	[spite] of all the
29	that in [spite] of the	15	in [spite] of his
27	[spite] of all	15	in [spite] of their

Table 4.9: The chains output for *spite* using the BBC corpus. Chains with a frequency of less than 15 have been omitted.

One can easily see that there is a lot of overlap, including a number of chains with identical frequencies, for example the two chains with a frequency of 17. These two are obviously from the same position in the corpus, and no information is gained from having both of them in the list. There is also a problem with upper and lower case forms, as in *in spite of this* and *In spite of this*. For the filtered version we normalise all words to lower case in addition to pruning using the cost function. We then get the following result:

895	in [spite] of
7	survive in [spite] of your parents
6	[spite] of all this
6	[spite] of government
5	[spite] of her failure to win an
4	[spite] of an official ban on strikes
4	[spite] of their recent confrontation with the
4	ahead in [spite] of her own request
4	gone ahead in [spite] of her own

Table 4.10: Chains for *spite* after pruning with the cost function

Here we have chosen a cut-off point of 4 in order to show more than just one chain; the actual filter would discard all but the first chain on the grounds that they are not frequent enough compared to the top one (using a threshold of ten percent of the highest frequency).

However, *in spite of* is an extreme case, where we are looking at an obvious multi-word unit, and that is reflected in the extreme frequency differential that we get here. In order to look at the full phraseology of *in spite of* we would need to re-classify it as a single word/token, and re-run the analysis to find larger units.

Other words that occur frequently in fixed phrases (such as *according to*, *because of*, *more than*) show similar results, whereas other 'normal' words do not. *Ship* (frequency 1389) has the two top chains *the ship* and *a ship*, and the overall frequency distribution is far less skewed:

490	the [ship]	41	[ship] has
183	a [ship]	39	cargo [ship]
71	[ship] was	39	indian [ship]
62	[ship] to	37	[ship] and
53	[ship] is	31	merchant [ship]
51	[ship] carrying	31	radio [ship]
43	[ship] in	30	ch [ship]
42	[ship] which		

Table 4.11: Chains for *ship* in the BBC corpus

So as a result of this step in the analysis we can not only describe the phraseology of words in general, but we can also identify with some confidence words that are candidates for multi-word units by looking at their frequency distribution. Without the cost function and the subsequent frequency filter we would retrieve a large number of overlapping and not very interesting chains. However, with the filtering mechanisms in place we can successfully discard irrelevant results.

4.3.3 Frames

A second method of extracting variable length multi-word units is based on collocational frameworks as described by Renouf and Sinclair (1991). Renouf and Sinclair defined those frameworks in advance and looked at the node words they found. In the work described here the procedure is reversed: starting from a node word, adjacent words are appended to the left and right hand sides if they are more frequent than the node. In order to allow for minor random variation, frequency bands (Quasthoff 1998, see above) are used instead of raw frequencies.

This procedure works well for content words, as they are comparatively rare; higher frequency words fail, as their adjacent words generally have lower frequency counts and are therefore not attached. For this reason the two methods have different strengths and weaknesses: frames are more linguistically defined, while *n*-grams work with any words regardless of their frequencies.

The view of syntax embodied in the 'frames' approach is that high-frequency words primarily act as 'glue', joining the more central (and less frequent) content words together, just as a wall is composed of bricks held together by mortar. When one disman-

tles a wall, the mortar usually clings to the bricks, and may still hold together some bricks. The 'chains' model does not presuppose the same view, as it does not make a difference between different types of words.

The following table lists the first 20 frames retrieved for *spite* from the BBC corpus. The 10 percent filter would leave only the top five in the list:

87	in [spite] of the	4	but in [spite] of
14	in [spite] of a	4	in [spite] of all the
14	in [spite] of this	4	in [spite] of an
10	in [spite] of its	3	and in [spite] of
9	in [spite] of his	3	but in [spite] of his
8	in [spite] of that	3	but in [spite] of these
6	but in [spite] of this	3	in [spite] of a record
6	in [spite] of their	3	in [spite] of all this
5	in [spite] of her failure to win an	3	in [spite] of an official ban on
5	in [spite] of these	3	in [spite] of government

Table 4.12: The twenty most frequent frames for *spite* in the BBC corpus

The main problem here is that *spite* is a fairly low frequency word, which means that many content words are more frequent and thus get added to the MWU in question, even though they should not really be considered part of it. But a frequency filter can remove those which are simply accidental, whereas frequently recurring ones remain.

Looking at the other example, *ship*, we get the following:

2	a [ship] at the lithuanian
2	a [ship] to the
2	after their [ship] was
2	demand to [ship] them in
2	of a [ship] off the west
2	of a [ship] which was
2	of the [ship] armed with
2	of the [ship] the
2	off the [ship] to
2	officials said their [ship] had to leave behind some five-thousand people
2	on the [ship] they arrived in
2	out if the [ship] is
2	reports from india say a [ship] carrying relief supplies for
2	that the [ship] is
2	that when a [ship] is
2	the [ship] had been
2	the [ship] has been
2	the [ship] is due to
2	the first [ship] carrying sri
2	the greenpeace [ship] gondwana has been

Table 4.13: The most frequent frames for *ship* in the BBC corpus

Here again we can see that there is no clear pattern emerging: all frames are equally frequent, and the even distribution indicates the lack of any frequently recurring units.

4.3.4 Synthesis

Both chains and frames perform similar functions, and have slightly different advantages and disadvantages, making it difficult to decide between the two. The obvious solution is to combine both of them: frames are capable of extracting additional candidates which chains would not cover (especially longer ones), and in the case of overlap (i.e. where the two methods identify the same candidates) those candidates would get an additional frequency boost.

The current approach to the identification of multi-word units can be split into two distinct steps:

1. retrieval of the candidate multi-word units

2. evaluation and ranking of the candidates

During the retrieval stage both methods are used to extract candidate multi-word units from the data. Each works differently and collects different kinds of units. Initially both methods were run separately, but it was found that combining the two sets of candidates gave practically the same results. Thus the outputs of both methods are combined to yield the complete set of candidates, which are evaluated in the second step of the analysis procedure. All candidates need to occur at least 5 times in order to count as recurrent enough. This avoids overly specific and rare sequences, though a value of 5 is of course arbitrary; anything greater than 1 would be suitable for excluding non-recurrent phrases.

The ranking or scoring of the candidates is then done by evaluating their frequency of occurrence and their length. The exact weight calculation is kept variable, so that different combinations can be explored. Apart from the cost function described above a simple weighting according to length is also provided. In fact, the cost function has been disabled by default, as the frequency boost through the frames candidates combined with simple length weighting is sufficient; instead the overall score for a candidate multi-word unit is calculated by multiplying its frequency by its length. Alternatively, the n^{th} power of the length (with *n* ranging from 2 to 4) can be used to bias the outcome towards longer units.

4.3.5 Problems

The current approach starts with a single word, and tries to identify multi-word units which include this word. This is in contrast to most other algorithms, especially those based on *n*-grams, which tend to operate on a complete text and try to find any recurrent word sequences in it.

The 'lexicographic' approach has a few additional problems, since we cannot be sure in advance what the phraseology of the word in question will be like. These problems tend to be less important if the most frequent *n*-grams of a *text* are extracted, as they are properties of the text, rather than properties of a lexical item:

- The word may not be involved in any multi-word units as defined above. It is conceivable that many words occur only in 'free structures'.
- The word may occur only in one or more multi-word units; however, it is not possible to know the number of units a priori.
- The word may occur both in free structures and in multi-word units.

These various possibilities make it difficult to evaluate the approach, which is basically exploratory, rather than an attempt to replicate any particular manual procedure. Its exploratory nature is a consequence of not following the traditional methods, which define multi-word units in terms of syntactic or semantic coherence. If a multi-word unit can be any combination of words, we cannot rule out some of the results as invalid without a full view of all units. If we can identify certain patterns in the result which give us some indication of the shape and form of multi-word units, we can then start

evaluating individual units according to criteria derived from those detected patterns. But we would then simply be using one exploratory method (clustering/pattern recognition) to evaluate another one (MWU identification), a dubious exercise unless the criteria can be justified in terms of well-known principles (such as Zipf's law).

While the lack of linguistic knowledge in this procedure makes evaluation very difficult, it also has distinct advantages. First, the detection of units does not depend on any information that could be based on *a priori* assumptions. The only assumption made here is that words can be combined in syntagmatic units that are re-used throughout language. Second, it is much easier to replicate the result of the study, as no further resources are required. And third, the procedure is basically language independent. There will be differences in applying it to other languages: for example, morphological variation in languages with a richer morphology will yield more distinct multi-word units, whereas a morphologically more simple language like English will feature fewer units and thus more repetition. But there is nothing inherent in the procedure that prevents it from being applied to data from other languages.

Regarding the possible outcomes, there is an important trade-off that needs to be considered: the frequency *vs* length of a multi-word unit. It is generally true that short sequences are more frequent than longer ones, especially when the longer sequence contains the shorter one. Each additional element in the sequence represents another choice point, and unless the unit is a fixed one that allows no variation and also does not permit its sub-sequences to occur without all elements, a choice point results in two or more longer units, which thus have a lower frequency than the shorter one they are derived from.

So frequency alone can not be used as a criterion to judge the value of a candidate

multi-word unit, as it decreases steadily with length. Length is not sufficient either, as otherwise the best multi-word units would be long but rare and thus not very useful. For this reason a weighted score has to be used to find a compromise between length and frequency.

4.3.6 First Conclusions

At this point we can assert the following conclusions:

- Both methods of identifying multi-word units (chains and frame) are able to identify valid units where intuition would predict them. This was demonstrated with one example only, but has been applied to other words as well (e.g. *front*, *sake* and *eye*) with positive results.

- There are slightly different outcomes depending on the procedure used, with implications for syntactic principles: the chains output corresponds more to the 'unit' view, where multi-word units are like words, only larger, and combine freely; whereas the frames results suggest a modular 'prefab' view, where utterances are made up of interlinked segments like a (one-dimensional) puzzle.

- For automatic processing/recognition of multi-word units it is possible to generate from the output of the identification procedures finite state automata for each multi-word unit, such as the one shown below (fig. 4.7). This would be suitable for use with the INTEX processing system (Silberztein, 1993) and could be used for large-scale retrieval of known units.

  This automaton would recognise the multi-word usages of *spite*. Interestingly, the unit *in spite of the fact that*, which was identified looking at *spite*, contains another

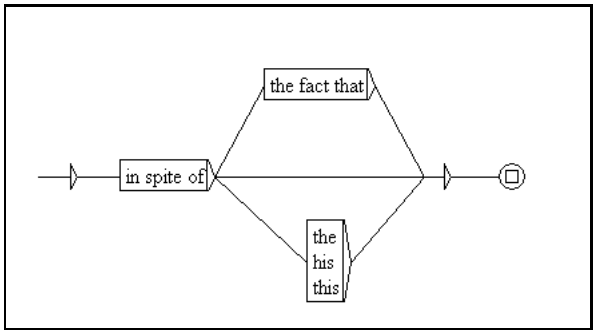

Figure 4.7: An INTEX-style automaton to recognise multi-word units related to *spite*

multi-word unit, *the fact that*, which is accounted for in the automaton.

The overall outcome of a comprehensive analysis of a corpus would then be a set of FSAs like the one in figure 4.7, which could be used to mark up multi-word units in texts efficiently, and is suitable for combination with automata describing other local grammatical phenomena (see Gross 1993, Mason and Hunston 2004).

4.3.7 Multi-word units as Grammar

The set resulting from applying the final version of the procedure to the 'obvious' non-unit *ship* is:

the ship, a ship, of the ship, the ship was, to ship, of a ship, ship in, ship was, ship in the, ship and

Most of these units are two words only, and apart from the first three the assigned score is distributed fairly evenly, which suggests no obvious bias (as would have been the case with the earlier examples). It would be easy to apply association measures

such as mutual information (see Church and Hanks 1989) in order to filter out such high frequency pairs.

However, the question arises whether such filtering is really desirable, or whether this result does not tell us more about the structure of language. In order to test this, a sentence was selected randomly from the Internet by searching for the sequence *ship in the*. The chosen sentence was *The Laird of Raasay, perceiving the ship in the harbour, went aboard to buy some wines and commodities*. For each word in this sentence the multi-word units were retrieved from the written part of the BNC to see whether there were any matches. The result of this experiment is rather surprising, as can be seen in figure 4.8.

The	Laird	of	Raasay	perceiving	the	ship	in	the	harbour	went	aboard
The	Laird	of									
		of	Raasay								
				perceiving	the						
					the	ship					
						ship	in	the			
							in	the	harbour		

to	buy	some	wines	and	other	commodities
to	buy					
			wines	and		
				and	other	
					other	commodities

Figure 4.8: Overlapping multi-word units computed for each word of a randomly selected sentence

This is a rather encouraging result: overlapping segments mirror larger phrasal elements (*the Laird of Raasay*, *perceiving the ship in the harbour*, and *wines and other commodities*), and places where there is no overlap reflect boundaries between segments. There are some parts which are not covered (*went aboard* and *some*), but the majority of

the sentence is accounted for. It suggests that this procedure can be used as a discovery procedure for larger units (see Harris 1955).

In fact, the analysis also resembles the overlapping grammar pattern analysis presented by Hunston and Francis (2000), who refer back to earlier work on linear grammars.

Further work would obviously be required to evaluate whether this approach to grammatical description works consistently on more example sentences, or even full texts. Language is essentially a one-dimensional, linear entity, and any hierarchical structure imposed on it will have problems. This is especially true for languages with free word order.

4.3.8 Multi-Word Units: Summary

The main aim of this section was to present a new way of identifying multi-word units, which starts from a word, rather than a text. This lexicographic approach has been shown to yield good results, though it is hard to evaluate due to a lack of objective criteria. In some ways traditional expectations are supported (chains), but it also seems that there are units which are not syntactically well-formed in themselves (frames) but suggest a different approach to analysing syntactic structure.

The problem boils down to the lack of an objective definition of a multi-word unit; but it is questionable whether a general definition can exist. There are certainly many different types of units, and one aspect that has not been addressed here is that of variability. Variations can include additional inserted words (such as *some* in the example sentence above) or substitutions (such as described by Renouf and Sinclair 1991). This

is more of an issue when looking at longer units.

There are a number of algorithms that have been developed to recognise multi-word units, and most of them find different types of units. The two algorithms presented here, chains and frames, recognise both independent and interlinked kinds of units.

Another issue concerns the move from the empirical exploration to a theoretical model based on the analysis, where we can use interlinked units to describe a syntactic sequence within a sentence. Further work is required in this area, but it should be feasible to develop a language model based on multi-word units, similar to the slot-and-filler model described by Sinclair (1991): large stretches of a sentence would be covered by overlapping multi-word units, while at some boundary points free variation is possible (up to a point). This needs to be confirmed on larger samples of text.

In this context it would also be useful to see what proportion of a text is made up of multi-word units. If we accept the general, interlinked definition it could be that only a small part is governed by free choice. This would support the Firthian view that language use is mainly routine, and does not contain much innovation (Stubbs, 1993).

The link between multi-word units and meaning seems central to their validity, as they embody the correlation between form and meaning, and provide a disambiguating context to the single lexical item. However, this does not always work, and one has to be aware that there are counter-examples, as in this excerpt from the BBC corpus: The frame in question is *of* **fire** *in the*; here the majority of lines are of the form *exchange(s) of fire in the [CRISIS REGION]*. But there are also two examples where *fire* is used in a different sense:

stressed the need for the speediest detection of [fire] in the cargo hold.

The other major concern is over the risk of [fire] in the tunnel .

While a few counter-examples do not invalidate the overall approach, there seems to be a second pattern at work, *[RAPIDNESS-ADJ] risk/detection of fire in the [CONFINED AREA]*, with fire being of the burning variety, rather than the military one.

Looking at a number of concordance lines for the sequence *detection of* in the same corpus, we can see that there is a more variable pattern emerging, which is roughly *earlier/speediest detection of [NEGATIVE ENTITY]*. As there is a lot of lexical variation, such patterns are still outside the grasp of automated detection. Similar observations can be made for *risk*, where most frames with the nominal sense contain *risk of*. With these usages of *fire* it seems that the disambiguating element is provided by the preceding pattern, as both *risk* and *detection* would only ever be used with the 'burning' sense of *fire*.

In conclusion we can say that this is a promising area of research, but it might be necessary to challenge traditional assumption about the structure of language. However, it is also very difficult to judge the outcomes, as with those assumptions gone there is no objective benchmark available for evaluation.

4.4 Lexis: Summary and Evaluation

In this chapter we have investigated three major elements of lexical information: lexical statistics, collocations, and phraseology. We have seen how we can extract useful information about lexical items from corpora, and how such information can be used

to build up a picture of the usage of a word of a kind that was not previously available. And the information gathered can be contrasted between different words in a corpus to get an idea of the structure of the vocabulary, or between the same word across different corpora to get an impression of differences between samples of language.

None of the methods requires human intervention, apart from the initial choice of parameters. Where possible, heuristics have been suggested to find suitable values. We have also stressed the necessity of making such choices explicit, as too many published studies are not repeatable due to incomplete information about their settings. And repeatability is an important aspect of empirical work.

However, the overall picture we are now able to create of a word is still incomplete, and in the following chapters we will now extend that in two directions: grammar and meaning. We will thus introduce further variables that can be used for partitioning both the set of word types and that of tokens for more detailed analysis. We could, for example, apply any of the lexical analyses described in the current chapter to a subset of tokens of a particular word type that occur in a certain grammatical environment. In that way our description will be able to reach a level of detail which would not be possible without this combined set of analytical procedures.

CHAPTER 5

GRAMMAR

5.1 Introduction

In this chapter we will look at the grammatical side of a word's behaviour. As in the previous chapter our approach is a lexicographic one: we start from a word, rather than a text. By applying a number of algorithms to the occurrences of a word form in the corpus we will try to explore how the word form is used in conjunction with other words. But this time we are not only interested in the co-occurrences with other words, but also in more abstract relations. Hence we will look not only at other word forms, but also at word classes.

In a previous chapter (see 2.3.2) we addressed issues with the traditional word class system. But while the present system is not without faults, it still serves as a useful layer of generalisation. As long as we are aware of the problems, we can still benefit from using it, substituting individual word forms or ad-hoc groups of words where a word class is too general. A further issue to bear in mind is that while the system of word

classes might be subject to change, phrases seem to be a lot more stable. So we need to be flexible, and not base too many assumptions on the word classes themselves.

In the process of analysing the grammatical environment of a word, we will make use of several kinds of programs that identify phrases through grammatical rules. This kind of analysis will have to be kept on a 'shallow' level with small grammars for two reasons: first, to stay as theory-neutral as possible, and second, to achieve a good degree of coverage and robustness of analysis. Decades of previous work in corpus linguistics (see for example Black *et al.* 1993, Sampson 1995) have shown how hard it is to syntactically analyse large amounts of authentic texts. The only major system that so far seems to be successful with large scale parsing is the constraint grammar developed at the University of Helsinki (Karlsson *et al.*, 1995).

With the phenomena we are investigating in this chapter we are still very much on the borderline between lexis and grammar, mainly because we are using the lexical item as the starting point for all of the processing. We need to distinguish this 'lexical grammar' from the Hallidayan 'lexico-grammar', as we do not regard lexis as the final choice in instantiating a grammatical construction: instead we view lexis and grammar as two intertwined areas which depend on each other through co-selection. Certain lexical items occur in certain syntactic patterns, and so certain syntactic patterns occur with certain lexical items. There is no hierarchy involved, none of the two areas dominates the other. They are merely different directions from which a single phenomenon can be approached. Incidentally, the same applies to the interface between syntax and semantics, which is a continuum rather than two clearly defined areas. This will be described in more detail in the next chapter.

For processing reasons it is easier to start from the word, as words can more easily

be retrieved from a corpus than structures. But once the complete set of words has been described, it should be easy to reverse the direction and investigate the structures that are shared between various lexical items.

Most work in computational linguistics in the area of syntax focuses on parsing, i.e. the analysis of sentence structure by machine using a (formal) grammar. The two most common approaches (on which more recent developments are based) are phrase structure grammar (as elaborated by Chomsky (1957)) and dependency grammar (after Tesnière (1959)), the latter becoming more common in recent times, since grammars based on phrase structure tend to get too complex as coverage increases. Dependency grammar parsers (e.g. Covington 1990) usually achieve good results with a few basic rules/principles and generally perform well in free word order languages where phrase structure grammar has difficulties.

This work is of little relevance to the empirical analysis of grammar, as it is simply trying to implement traditional grammatical formalisms with the aim of testing and further development. One important reason is coverage: formal grammars are not robust enough to cope with authentic data, and often work only on test sets of artificially created sentences. Realistic data requires large grammars. Where corpus data has been processed syntactically (e.g. the Penn treebank, Marcus *et al.* 1993, or Sampson's SUSANNE corpus, Sampson 1995), it was mainly done manually, possibly with the aid of shallow parsers which attempt only a partial analysis. Manning and Schütze (2000, 414) report that *[t]he treebanking manual for the Penn Treebank runs to over 300 pages*. Sampson (1995), the description of the analytic scheme of the SUSANNE corpus, is 500 pages long.

Shallow parsing (which does not attempt a detailed analysis, but rather concentrates

on the basic structure of a sentence) is relevant in that a partial analysis on a theory-neutral level can be used as a starting point for further automatic exploration. As mentioned above, certain traditional categories can usefully be employed for this purpose, and a shallow parser would typically identify phrases without attempting to work out the complete structure of a sentence (which is where most problems are introduced).

The other important strand in the computational processing of syntax is taking the next logical step from shallow parsing: determining the most likely attachments for prepositional phrases (Hindle and Rooth, 1993), verbal subcategorisation frames (Briscoe and Carroll, 1995), and basic syntactic patterns (e.g. Brent 1993). These will in part be covered by section 5.3 on usage patterns.

We will now look at three different aspects of the grammar of a word: first, *colligation*, an extension of collocation which focuses on general categories rather than lexical items; second, *usage patterns*, which are somewhat related to collocations, but describe syntactical dependencies rather than simply spatial ones; and thirdly, *grammar patterns*, which try to capture the typical grammatical environments of a word in a (finite) number of common patterns.

5.2 Colligation

As Hunston (2001) observes, the term *colligation* has not been used much since its original introduction to linguistics by Firth (1957). The related field of collocation has attracted more attention in corpus-based work, presumably because it is a concept that can be defined more easily and also fits better into the traditional division into lexis and grammar unlike colligation, which is located somewhat between the two. Willis (1993)

talks of the *grammar of class* when referring to the tendency of particular nouns (of a semantically defined class) to co-occur with the same delexicalised verb. He used the term *collocation* for describing similar co-occurences, which might be better described as colligations.

Hoey (2003) interprets colligation in a slightly different way, applying it to larger structures within a text, such as the preference to be part of the *theme* or occurrence in certain positions within elements of a text. As we are not dealing with individual texts here, but a continuous stream of utterances, we will not follow his usage.

Earlier, and with reference to lexical items in general, Hoey (1998) lists three aspects of colligation. The first one (in analogy to Firth's famous definition of collocation) is *the* **grammatical** *company a word keeps (or avoids keeping)* (my emphasis). The question then is what exactly 'grammatical' means here. Hunston (2001) interprets it as referring to the grammar patterns (which we will look at in section 5.4 below). The second aspect concerns the grammatical functions in which a word occurs frequently (or not at all). Hunston here refers to an observation by Francis (1991) that certain words tend to occur in a limited range of clausal elements, e.g. predominantly as adjuncts or nouns. We will deal with this in more detail in section 5.3. Hoey's final aspect concerns the preferred place a word takes (or avoids) in a sequence. This would relate to multi-word units as discussed previously (see 4.3).

In the words of Stubbs (2001, 65), *[c]olligation is the relation between a pair of grammatical categories or, in a slightly wider sense, a pairing of lexis and grammar*. Stubbs also points to Francis (1993) for an investigation into colligational behaviour. Overall it is difficult to define exactly what is meant by colligation, apart from the fact that it is distinguished from collocation by the use of grammatical categories rather than simply

lexical items. In Mason (2000b) colligation was (for the purposes of lexical gravity computation) interpreted as replacing word forms with their respective parts-of-speech labels, but ideally we would want to include larger categories (e.g. phrasal elements) as well.

The question that we want to answer by looking at colligational patterns is in what grammatical contexts does an item occur. Francis (1993) gives the example of *as ADJ/ADV as possible*. Here part of the information gained from the corpus description is that *possible* can frequently be found with this structure, whereas most other adjectives cannot.

This makes colligation similar to the phraseological patterns investigated earlier (in section 4.3): but now we are not using exclusively lexical items. Instead we employ a mixture of lexical items, parts-of-speech labels, and phrasal categories, since we cannot be certain of the restrictions that apply for any given structure. If we were using lexical items only, we would not be able to pick up the above structure as there is too much variation in the 'ADJ/ADV' position, so we would find only the most frequent ones without realising the underlying pattern. Using parts-of-speech labels only we would miss the *as ... as* part, and would think that any conjunction or particle could be used in those positions. And though not applicable to this example, phrasal categories allow us to generalise by not having to worry whether an adjective is used in a noun group or not, again permitting some variation within pattern components.

In order to implement a colligation recogniser we have to extend the phraseological 'chains' procedure (see section 4.3.2) to include intermediate categories. For that purpose we will use a chart parser to process the instances of the words as we find them, and then we will extract all possible paths in the chart which include the node word.

Again we will use paths of varying length, in analogy to the n-grams of different sizes.

As the phrase structure grammar used for the chart parser (see figure 5.1) introduces a large number of intermediate constituents which we are not interested in, we will try to filter those out as we go along. We also need to restrict the length of the n-grams collected, as otherwise their number will be too large for processing. In addition to the lexical items collected, there will be several alternative (phrasal) links between chart positions, and consequently a large combinatorial explosion of n-grams.

```
v       →  VB|VBZ|VBG|VBD|VBN
prep    →  IN
adj     →  J*
adv     →  RB*
det     →  DT
pron    →  P*
N1      →  N*
N1      →  J* N1
conj    →  CC
nounP   →  DT N1
nounP   →  N1
nounP   →  nounP "of" nounP
prepP   →  prep nounP
verbP   →  v nounP
verbP   →  v
to-inf  →  "to" VB
clause  →  nounP verbP
clause  →  nounP verbP nounP
that    →  "that" nounP verbP
```

Figure 5.1: The context-free phrase structure grammar used for identifying constituents for processing colligation

One aspect that we cannot describe here is that of semantic categories such as 'negation', which are sometimes grouped with colligation; these are more akin to discourse prosodies, which are currently outside the scope of automatic recognition as argued in section 2.3.5. It might be possible, in principle, to use a few basic heuristics to identify

negation, such as the occurrence of a set of lexical items (*no*, *not*) or morphemes (*in-*, *un-*). However, we have not pursued this for the current project.

5.2.1 Computing Colligation

For the purpose of colligation we are looking at *n*-grams of a length between 3 and 5; they are collected in the same way as the 'chains' described earlier. The main difference is the inclusion of word classes and phrasal categories together with lexical items.

5.2.2 Colligation Examples

Table 5.1 lists the most frequent colligations of the word *mine* from the BBC corpus. Many of the occurrences seem to be related to the nominal uses, but the most frequent one is the possessive pronoun. With this we can see that it often follows as a qualifier to a noun phrase, e.g. as *a colleague/the actions/friends of mine*. An interesting counter-example is *a group of mine-workers*, where the hyphen was interpreted as a token boundary.

5.2.3 Evaluation of Colligation

Colligation, despite being about half a century old, is still a largely unexplored concept. Only recently have researchers begun investigating it, as for example Hoey (1998) or Hunston (2001). The operationalisation presented here might be a first step towards a systematic description, but it clearly needs more work in order to facilitate interpretation by a human analyst. One further step might be to see what the coverage of the colligation patterns of a word would be within a text, in other words, how typical they

76	nounP *of mine*	16	det adj *mine*
66	det nounP *mine*	15	det *mine* prep
44	*a* nounP *mine*	15	nounP *mine* v
40	prep nounP *mine*	15	prep *mine* nounP
35	*mine* prep nounP	14	*mine* nounP .
34	nounP *mine* nounP	14	*mine* nounP prep
34	nounP nounP *mine*	14	*mine* prep *the* nounP
32	nounP prep det *mine*	14	nounP *iron ore mine*
32	nounP v *mine*	14	nounP *iron* nounP *mine*
30	nounP *mine in*	14	nounP v *ore mine*
30	nounP *mine* adv	13	*mine in* det nounP
29	*mine* v nounP	13	*mine* nounP nounP
28	*mine in* nounP	13	*mine* nounP prep nounP
28	nounP *mine* adj	13	*mine* prep v
27	nounP prep *the mine*	13	*the mine* nounP
24	nounP v nounP *mine*	13	adj *mine* nounP
23	*mine* . nounP	13	nounP *'s* nounP *mine*
23	nounP *of* det *mine*	13	nounP pron nounP *mine*
22	nounP *at* nounP *mine*	12	*a land mine*
22	nounP *mine* prep	12	det *land mine*
21	adj nounP *mine*	12	nounP *mine (*
20	*mine* prep det nounP	12	nounP prep det *anglo-american mine*
19	nounP *mine .*	11	nounP *at a gold mine*
19	nounP *of the mine*	11	nounP *at* det *anglo-american mine*
18	det *mine* v	11	nounP *at* det *gold mine*
18	nounP *mine near*	11	nounP prep *a gold mine*
18	v det *mine*	11	nounP prep det *gold mine*
17	*the* nounP *mine*	11	v prep *mine*
17	det *mine* nounP	10	*at* det *mine*
17	nounP *mine ,*	10	*mine* nounP *and*

Table 5.1: The colligations of *mine* (BBC corpus)

are as a reflection of the word's grammatical behaviour. That, however, goes beyond the scope of the current project.

5.3 Usage Patterns

By usage patterns we mean pairs of lexical items which are related to each other in a syntactic relationship. They are *patterns*, since they are varying instances of the same relationship, and they reflect *usage*, since they record the number of times a certain combination has been used in the corpus. Owing to the non-random nature of language we can assume that any existing patterns in the word combinations will emerge from the collected data.

In the next chapter we will exploit this feature in order to investigate word meaning, but before that we will study how the analysis of usage patterns can benefit our understanding of grammatical regularities. The next section will introduce the notion of usage patterns and give the rationale for their use in analysis; then we will look at the inventory of patterns analysed and at issues in the identification of usage patterns, and then we will describe how the gathered data can be evaluated for the purpose of description. Finally we will summarise the outcome of this case study.

5.3.1 Introduction and Rationale

In traditional approaches to grammar (see for example Sells (1985)) constraints were soon discovered which limit which nouns can occur as the objects of particular verbs. These selectional restrictions can sometimes be expressed via semantic groupings, but mostly they have to be lexicalised, as there are no easily identifiable regularities; instead many verb-object relations seem to be idiosyncratic, i.e. they cannot be described by rules. This leads to the relegation of such subcategorisation information from the grammar into the lexicon.

For practical applications such as the attachment of prepositional phrases researchers (e.g. Hindle and Rooth 1993, Brill and Resnik 1994, Pantel and Lin 2000) have gathered distributional information to work out the probabilities of a particular noun within a prepositional phrase being either a postmodifier of a preceding noun phrase or an adjunct to the corresponding verb. Certain verb/noun and noun/noun combinations are more likely to be verb-noun-adjunct than verb-noun-qualifier, and knowledge of the distributional properties can help make the right decision in determining the structure of the sentence.

Such information can also be valuable when describing the tendencies of words to occupy particular positions within a sentence. Francis (1991) describes how *lap* predominantly occurs in adjunct position. It can be expected that there will be a bias for most words to behave in a similar special way, if only because certain kinds of nouns denote entities/concepts that cannot perform actions themselves and thus would rarely be in subject position.

Apart from clause positions, other information about words might also be relevant to a description of their typical usages. The use of usage patterns here is similar to word sketches (Kilgarriff and Tugwell, 2001); commonly used patterns will emerge through their higher frequency. We have chosen a number of other word relations which can easily be recognised by the computer without introducing too many errors.

Studies based on clause relations have been done before, e.g. by Hindle and Rooth (1993), but it seems that they rarely go beyond case studies, and that they are not used for a comprehensive description of language. In this project, however, we will do exactly that, integrating usage patterns into the resulting language description.

We will first describe the pattern inventory, the list of patterns that the recogniser can identify. Then we will show how they are recognised in the text, and with what success rate. We then discuss some typical errors or problematic issues in the recognition process, before describing how the information gathered can be used for our purposes.

5.3.2 Pattern Inventory

Without a detailed analysis we will not be able to capture all interesting syntactic relations between pairs of words. But it is not possible to do a detailed analysis of unre-

stricted text in a fully automated way, so we have to compromise by restricting ourselves to a shallow analysis. That means that some usage patterns remain outside the scope of our analysis. This is the main problem of using usage patterns for descriptive purposes, as we shall see below.

The following patterns are recognised by the analyser:

AN adjective-noun, for example *psychological distinction*. This is a simple modification.

NN noun-noun, for example *division walls*. These are compound nouns or nouns used as noun modifiers; sometimes the first noun can also be classed as an adjective, as in *manual systems*. This relation also includes 'noun *of* noun', so that for example *bottle of wine* is treated the same as *wine bottle*.

Ninf noun-infinitive, for example *parliament give*. This would be an infinitive clause where the noun acts as a subject, such as *the decision for parliament to give ...* [I]

PN preposition-noun, for example *with effects*.

SV subject-verb, for example *liners are*, or *opponents prepared*. This would be the first noun phrase before a verb; complications arise with the passive voice.

VO verb-object, for example *consider position*. No distinction is made between objects and complements; this is always the first noun phrase following a verb.

VP verb-preposition, for example *exhibited in*, or *dominated by*. No attempt is made to distinguish between adjuncts and noun phrase qualifiers, as prepositional attachment is still an unsolved problem.

Vinf verb-infinitive, for example *designed (to) take*, or *trying (to) circumvent*. This is an infinitive clause as a complement to a verb.

In all these cases we are storing phrase heads only, as they generally seem to be the most salient word in the phrase. This obviously does not apply to the patterns where the individual elements are not clauses, such as `AN` and `NN`.

5.3.3 Pattern Identification

Unlike most other procedures described in this project, usage patterns are identified by processing the full corpus, rather than a set of concordance lines. The text is tokenised and split into sentences using a recogniser that evaluates possible sentence-final punctuation and decides whether a full stop does in fact mark the end of a sentence. The words in each sentence are then tagged for word classes and processed by a chunker which will recognise possible phrases. A simple finite state grammar is employed to identify noun phrases and verb phrases; the two automata used are depicted in figures 5.2 and 5.3.

One basic problem with this approach is that it is limited to a particular view of phrase relations. It presupposes that phrases are the elements which enter grammatical relations, whereas authentic language is far more complex. This is obvious to anybody who has tried to teach students the SPOCA type of syntactic analysis, which is similarly geared towards phrases. Here the basic clause elements are assigned roles (Subject/Predicator/Object/Complement/Adjunct) based on their function within a clause. Schemes like SPOCA only work well with simple clauses, which hardly occur in real language. The main idea is that clauses have a 'logical' structure, describing entities and their relations. However, we often find that in authentic language other, embedded, clauses take on the role of object or complement, and these clauses in turn have their own internal structure. Basic SPOCA is far too simplistic to be able to describe such

State	Arc	Target
0	JJ* PP$ PN CD OD	1
0	N*	2
0	DT PDT	4
0	PP PN PDT DT	5
1	JJ* OD CD VBG	1
1	N*	2
1	CC	3
2	‘of’	0
2	N*	2
2	POS	1
3	JJ*	1
4	JJ* CD OD	1
4	N*	2
4	RB	6
5	‘of’	0
6	JJ*	1

Figure 5.2: The automaton for noun phrases. The initial state is 0, and terminal states are 2 and 5. The word class labels are listed in appendix A

structures.

As the usage pattern recogniser is ignorant of clause boundaries (it operates on the sentence level, and tries to identify predicators as core elements of a clause) it will run into problems with complex structures. Here are a few example sentences and the identified relations:

I think we have a difficult problem.		
AN	difficult	problem
SV	i	think
SV	we	have
VO	have	problem

State	Arc	Target
0	RB*	0
0	MD DO*	1
0	HVZ HVD HV MD	2
0	BEN BEZ BED BEDZ BER	3
0	VBD VBZ HVZ BEZ BED HV VB DO BEM BER BEDZ HVD VBN	5
1	RB*	1
1	HVZ HVD	2
1	VBD VBZ VB VBN	5
2	RB*	2
2	BEN BE	3
2	VBD VBZ VBN VB	5
3	RB*	3
3	BEG	4
3	VBD VBZ VBN VB VBG	5
4	RB*	4
4	VBG VBD VBZ VBN	5
5	RB* RP	5
5	XNOT	4

Figure 5.3: The automaton for verb phrases. The initial state is 0, and terminal states are 3 and 5. The word class labels are listed in appendix A

The second clause which is the object of *think* is not recognised as such. The other relevant relations have been successfully identified. The problem arises from the purely word-based approach, as clausal elements are not recognised at all. Given this restriction, the analysis has been successful; it avoided the invalid relation `VO think we`.

Unlike most other procedures described in this project, usage patterns are identified by processing the full corpus, rather than a set of concordance lines.

AN	full	corpus
NN	other	procedures
NN	usage	patterns
NN	concordance	lines
PN	in	this
SV	procedures	described
SV	patterns	identified
VP	described	in

Other has been wrongly identified as a noun. There is also a problem with the qualifier *described in this project*, where the passive is not recognised. The preposition-noun relationship should have taken *project* as head, whereas it took *this*.

It has to be said that the overall quality of the recognition is not very satisfactory. This is partly due to the complex nature of language, and partly to the quality of the recogniser itself. It should be possible, given further resources, to extend the grammar, and also the heuristics used for the recognition of patterns, in a way that would lead to improved results.

However, since the analysis of usage patterns is only a small part of this project, we will try to evaluate what useful information can be gained from the recogniser in its current state. The result, it is hoped, will enable us to judge whether this is a way into grammar that is worth pursuing.

5.3.3.1 Quality of Analysis

Evaluating the quality of the recogniser can be expressed it in terms of *precision* and *recall*: precision gives the ratio of correctly identified relations to incorrect ones, and recall gives the ratio of recognised relations to non-recognised ones. For this evaluation we chose a sample text and processed it with the usage pattern recogniser in debug-mode. We got each sentence followed by the recognised usage patterns, and we could then compare the output with the analysis of a human analyst (in this case the author himself).

The sample text analysed was a random excerpt from the FLOB corpus, of 596 words in length. The recogniser identified a total of 181 relations in the text. The correctness of the relations was then checked manually, incorrectly identified relations were marked up and missing relations were added. The evaluation procedure was based on the capabilities of the recogniser: in other words some relations that had not been recognised were not marked as such if the recogniser could not have possibly identified them; one such case is the coordination of subjects or objects.

Up to a point this is a questionable decision, as in fact the analysis is less comprehensive than the result seems to suggest. But we can hardly evaluate a shallow analysis by taking a deep analysis as a benchmark, unless we want to see how far we can get with the shallow procedure. Here we were interested only in the success rate of the recogniser itself, i.e. how well it did the job it was designed for. We were fully aware of its limitations, so there was no need to complicate the evaluation any further.

When we inspected the errors, one type of wrongly identified relations stood out, namely possessive pronouns and nouns, as in *their king*, which were recognised as `NN`,

when there was in fact no suitable category. As this is not a grave error we have decided to perform two calculations, one with those relations marked as errors, and one with them marked as correct. The text with all relevant relations is included in appendix E.

The evaluation resulted in 181 relations, of which 142 were correctly identified. This leads to a recall value of 78.5%; taking the pronoun-noun relations as correct we have 153 correctly identified relations and a recall of 84.5%. The recogniser identified 184 relations in the text, of which 142 were correct, giving a precision value of 77.2%, or 83.2% when counting the pronoun errors (11) as correct.

Given the minimal effort invested in putting together the recogniser these results are excellent, and clearly sufficient for our purposes. Most of the errors can be traced back to either the part-of-speech tagger or the grammar, which is really too small to capture many of the intricacies of noun phrases using non-nominal categories. The results could also indicate the inadequacy of the word class system.

There are also a few processing errors where words have been put in single quotes; these errors could have been avoided by stripping any 'superfluous' punctuation, but then other errors might have been introduced. Often a single error is counted multiple times, because a recognised relation is wrong and therefore the correct relation is missing. From the analysis it also became apparent how many clausal objects occur in the text sample, whereas the recogniser is limited to lexicalised objects only, as mentioned above.

An inspection of the recogniser output has further highlighted two major issues, which we will now discuss in more detail.

5.3.3.2 "The Passive Voice Should Be Avoided"

One serious problem concerns the analysis of sentences in the passive voice. In the (attested) example *Often infected people are rejected by family and friends*, the subject-verb relation is `people rejected`, but it is not really the people that do the rejecting, instead *people* is the logical object of the verb, rather than its subject. The reversal of the subject/object relation (and the relegation of the subject into an (optional) adjunct) poses a fundamental question: should we take the syntactic relation as it is, or should we apply a transformation, normalise the sentence structure in order to extract the logical subject and object, rather than the grammatical ones?

In the end it depends on the purpose of the analysis. If we want to find out which nouns are used as subjects of a particular verb, then the logical subject would make more sense. We can always recover the information that the verb has been used in the passive voice, as we are collecting that separately (see section 4.1.4). The only real problem is when we want to count how often a noun occurs in subject or object position, in which case the statistics get confused when applying different rules to passive clauses. However, in a subsequent step we will process the extracted syntactic relations to identify semantic relations; and that would require the logical rather than the purely grammatical relations.

This grammatical-logical mismatch introduces a dilemma with implications for the empirical status of the analysis. We have to introduce additional preconceptions, knowledge about the structural differences between active and passive sentences which we did not gain from the analysis itself. While it may be acceptable to allow this extra processing step with active/passive sentences, where do we draw the line with other

phenomena? Our existing knowledge of language may give us grounds to argue that this is indeed a special case which needs special treatment, but it still 'pollutes' the purity of the empirical approach.

In the end we could always justify our solution by referring to systemic functional grammar, which has the categories 'agent' and 'patient' going beyond the grammatical realisation as subject or object. It then becomes a matter of the level of description, whether we want to talk about the (grammatical) form (subject/object) or (logical) function (agent/patient).

5.3.3.3 Local vs Global dependencies

The usage pattern recogniser has only a limited window within which phrases are analysed for their potential relations. This sometimes leads to the identification of relations which do not really apply. An example for this is the clause *I should have gone to Constantinople to learn Arabick* from the corpus of 19th Century novels. Here the recogniser identifies *learn* as an infinitive complement of the noun *Constantinople*, in analogy to the valid (invented) example *I had a good reason to learn Arabic*. In the latter clause the infinitive does qualify the noun, whereas in the corpus example it acts as an infinitival adjunct, dependent on *gone*.

This is a real problem with any kind of shallow analysis, which only takes into account surface features in a localised context. It also highlights the fact that the traditional word class system may not be all that suitable for this kind of processing, as identical surface structures clearly have different underlying structures, in this case relating to the attachment of the infinitive clause. Perhaps there ought to be a subclass of `noun` which typically takes an infinitive complement; or that should be made an explicit

feature of nouns such as *reason*, in which case the grammatical analysis needs to operate with a mixture of lexical items and word class labels in order to analyse constructions where the surface structure described in the traditional word class system is not sufficient. As an aside, this is a general problem when treating syntax as independent of other areas of linguistics.

However, in terms of the quantitative analysis we can assume that it will not affect the overall result too much. Relations such as `Ninf Constantinople learn` will remain accidental, whereas `Ninf reason learn` will be repeated multiple times. We could also investigate how often a particular noun occurs with an infinitive complement to cater for a variety of verbs. We would then discover that *Constantinople* is rarely followed by an infinitive, whereas *reason* often is. These statistics can be used to filter out random 'noise' created by mistakes in the analysis, similar to the research into PP-attachment mentioned above. In fact, in order to avoid such purely accidental relations, only those which occur more than once in the corpus are stored; single cases are filtered out on the grounds that they would not contribute anything statistically significant anyway.

8 much, me
7 him
6 nothing
3 you, way, trouble, time, them, curiosity
2 world, wish, stamp, something, science, school, reader, lesson, inclination, himself, disposition, desire, boys, body, anxiety

Table 5.2: `Ninf` usage patterns with *learn* as the infinitive

Table 5.2 shows all the `Ninf` relations where learn is the infinitive, sorted in order of frequency. As expected, *Constantinople* has been discarded as a singular occurrence, but there are several aspects of this list which need further explanation. It seems that there

are several underlying patterns to this surface structure:

- as a qualifier of a preceding quantifier (as in *much to learn*, or *nothing to learn*),
- as a qualifier of a preceding noun specifying feelings related to the process (*trouble to learn*, *curiosity to learn*),
- as a subject/verb relationship in a subordinate clause that is realised through an infinitive (see *me to learn*),
- as an object/verb relationship in a subordinate clause that is realised through an infinitive (see *science to learn*),
- as a misinterpretation due to an enclosed adverbial phrase (*paid so much to learn*, which is really `Vinf paid learn` instead).

```
h phrases got by heart , With much to [learn] and nothing to impart , The youth obe
 clipper ways , but he hain't much to [learn] . Steer he can - no boy better , ef I
 even in those days there was much to [learn] from him ; and above all his fine spi
 hich shows that I have still much to [learn] . " " I fancy it 's some local practitio
 , " said he . " We have both much to [learn] , and we shall both be better men for
  to such a work . That I had much to [learn] , myself , before I could teach others
  my poor father wished me so much to [learn] , and although I am so anxious to lea
ion ; and I , who had paid so much to [learn] the beginning , might pay a little mo

spered . " Yes . They have sent me to [learn] what had befallen you . " " They discov
 sbon . It is very important to me to [learn] how Wellington 's troops are distribu
honi soit ? Ah , it is hard for me to [learn] , hard for me to dare to be myself .
 Mahbub Ali should have come to me to [learn] a little lying . Every time before t
or the Appin Stewarts , enabled me to [learn] , and helped me to understand , about
lieve there is nothing left for me to [learn] . I presume I may say that I know al
olish governess , do you expect me to [learn] my lessons , when I haven't got you t
ducation ; so that if he wished me to [learn] , he should rejoice at my misfortune .

 or we would not be at the trouble to [learn] a language , if we could have all tha
should not take the slight trouble to [learn] how to make it heard is one of the s

o be sure , if a man has a science to [learn] , he must regularly and resolutely ad
```

Table 5.3: A selection of relevant concordance lines for *learn* as an infinitive

Here we have the same problem already mentioned in section 2.3.1, in that the same surface structure has a variety of interpretations. But it appears that this dilemma can be

more easily resolved, since the underlying structures correlate with the lexical choices. Object pronouns or nouns referring to people indicate the subordinated subject/verb relation, while nouns expressing feelings or emotions indicate the stance towards the process described by the infinitive. We therefore have another example of grammar and lexis being interdependent. By taking both into account we can categorise the data in a satisfactory way, which would not be possible by syntax alone.

5.3.4 Evaluating Usage Patterns

Despite the problems dealt with in the previous sections, we can make good use of the information collected through the recognition of usage patterns. Several kinds of descriptive information can be collected directly from the list of relations, and further processing of semantic information is described in the following chapter (see section 6.3):

Grammatical distributions: how often does a word occur in a certain position within the clause?

Syntactic arguments: what other words occur in a particular syntactic relation with our target word?

5.3.4.1 Grammatical Distribution

The following example, FIRE, has been taken from the BBC corpus. It has been selected mostly randomly, without knowing in advance what the result would look like. Initially we planned to look at LAP as well, but most occurrences in the BBC corpus are simply in lists of racing results, which are not suitable for this kind of analysis.

5.3.4.2 FIRE

The test word used for evaluation was the lemma FIRE with all its inflected forms. It was chosen because of a noun/verb ambiguity, and the output of the usage pattern processor is as shown in table 5.4. The first number in each cell refers to the first position in the pattern, e.g. **A** in **AN**, the second number to the second position. Since *fire* cannot be an adjective, the first number in the relevant cell is zero.

Relation	fire	fires	fired	firing
freq	4147	316	1296	574
AN	0/178	0/13	0/0	0/31
NN	193/449	12/64	0/0	35/14
Ninf	23/34	0/0	0/0	0/0
PN	0/539	0/48	0/0	0/57
SV	264/25	43/7	0/565	45/18
VO	107/1557	0/88	561/0	11/35
VP	320/0	8/0	837/0	19/0
Vinf	0/43	0/0	2/22	0/0

Table 5.4: Usage pattern distribution across inflected forms of FIRE

The figures given in the table are token frequencies, i.e. multiple occurrences of the same usage pattern instance are counted separately. The discrepancies between the overall frequency of the word form and the number of occurrences in the usage pattern counts can be explained by the fact that not all instances of a word are recognised as a usage pattern; for example, if no subject can be found in an instance of *fire* as a verb, it will not be listed. This somewhat reduces the scope for interpretation, as we cannot be certain of the total numbers. However, the recall figures reported earlier indicate that this is not a significant problem. Intransitive usages, on the other hand, are a problem; in this case the verb appears only in the `SV` relation, provided a subject can be identified.

We could say from table 5.4 that *fires* is more frequently used as an object than as a

subject, that *firing* is used as a nominal and a prenominal modifier, but never followed by an infinitive, and that *fire* is often used in a prepositional phrase. One surprising observation is that *fired*, a finite form, is recognised as an infinitive complement of a verb. The actual word form is in fact slightly misleading, as an investigation of this result reveals. Here are some sample concordance lines:

```
 ORNO TROOPS FIRE Soviet troops are [reported] to have fired into the air over the
  area . Government forces are also [reported] to have fired a missile into the Sa
 armed with automatic weapons , are [reported] to fired at random on a group of wa
. ) ZIMBABWE POLICE Riot police are [reported] to have fired warning shots over th
yal to General Holomisa . Shots are [reported] to have been fired into the buildin
 t early today when the Iraqis were [reported] to have fired surface to surface mi
 were killed . Security forces were [reported] to have fired on the demonstrators
out their vehicles . The police are [reported] to have fired shots to break up the
 ists trying to get to the town was [reported] to have been fired upon , but this h
ng , and said several missiles were [reported] to have been fired at the town from
```

So the feature ‘non-finite’ is carried by the auxiliary, apart from the third line, which seems to be ‘non-grammatical’.

5.3.4.3 Syntactic Arguments

The syntactic arguments are the word forms that co-occur with a target word in a particular usage pattern. Conceptually they can be viewed as similar to collocations, with a different definition of the environment: instead of a purely spatial approach, a range of words on either side, the environment is defined syntactically through the relation in question. We can then apply similar methods to extract ‘syntactic collocates’ from the set of syntactic arguments. However, frequency alone tends to give a good picture of how a word is used.

As an example we will now look at the AN relation of the inflected forms of FIRE. These are given in tables 5.5 to 5.7.

A	N	freq
heavy	fire	42
friendly	fire	18
automatic	fire	17
serious	fire	12
huge	fire	9
reported	fire	8
indiscriminate	fire	7
small	fire	7
sporadic	fire	7
intense	fire	6
big	fire	6
biggest	fire	5
fierce	fire	4
industrial	fire	3
massive	fire	3
own	fire	3
anti-aircraft	fire	3
major	fire	2
nuclear	fire	2
sacred	fire	2
severe	fire	2
busy	fire	2
famous	fire	2
fatal	fire	2
subsequent	fire	2
universal	fire	2

Table 5.5: adjectives modifying *fire*

A	N	freq
big	fires	3
serious	fires	3
small	fires	3
frequent	fires	2
smaller	fires	2

Table 5.6: adjectives modifying *fires*

A	N	freq
heavy	firing	17
indiscriminate	firing	5
sporadic	firing	4
intermittent	firing	3
long-range	firing	2

Table 5.7: adjectives modifying *firing*

Obviously, *fired* does not occur in this relation, as it cannot be a noun, though in theory it could have appeared as the first element, the adjectival modifier.

What we can see from these tables is that *fire* in the BBC corpus mainly refers to shooting (*heavy*, *friendly*, *indiscriminate*), and occasionally also to flames (*huge*, *massive*). Some instances are ambiguous (*serious*, *intense*). The plural *fires*, however, refers only to flames, and is modified only by adjectives indicating size or severity. *Firing*, on the other hand, is military again. This is another clear example of the correlation between form and meaning. Here it is actually the uncountable noun which has the military meaning, whereas flame-fires are countable. The nominalisation through the *ing*-form of the verb is also restricted to the former meaning.

5.3.5 Usage Patterns: Conclusion

In this section we have investigated the usefulness of usage patterns to describe grammatical regularities. While we can identify the set of patterns from the chosen inventory with both high precision and recall, the coverage overall is not so good, as the usage patterns are fairly simplistic and not suitable to describe more complex grammatical structures. So recall within the set of recognisable relations is satisfactory, whereas

overall recall in terms of grammatical relations in a given text is not.

This is partly a problem of the shallow approach used here to identify patterns. A more comprehensive syntactic analysis might be able to handle more complicated sentence structures, but would require a much greater effort in the development stage than the 'quick and dirty' heuristic analyser presented here. And comprehensive broad coverage parsers that are sufficiently robust to work with unlimited data and without human supervision are notoriously difficult to develop.

Altogether the information collected here is too sparse to be really useful for a full syntactic description. However, as we will see in the following chapter, it can still be usefully exploited to derive information about the structure of the vocabulary, which aids the semantic description of language. Here we are assigning preference information of the kind *what is a typical subject of the verb xyz?* to the area of semantics, although it is really on the borderline between syntax and semantics.

5.4 Grammar Patterns

In principle, no sharp division between lexis and syntax exists. Most corpus-based research so far has shown this (starting with e.g. Sinclair and Jones 1974). Instead we have a continuum between fixed phrases, longer prefabricated multi-word units (see Danielsson 2001) on the one hand, and lexically variable, syntax-driven constructions on the other. Complete variability is not likely considering what we know about the principal mechanisms of language, so the 'slot-and-filler' model describes only theoretical possibilities (explored in great detail by intuition-based linguistic research). In principle those possibilities put no constraints on co-occurrence, but the reality of 'productive'

syntax as evidenced by usage (see the previous section) remains much more limited. Due to lack of adequate data analysis in grammar, theoretical linguists overrate by far the potential of the slot-and-filler model.

The paradigm followed here results from an amalgamation of different approaches with the following properties:

- **local** — they describe only very localised phenomena and sentence fragments/phrase, thus avoiding unexpected 'side-effects' of rules in other contexts
- **lexical** — they operate on a mixture of category labels and lexical items, thereby avoiding problems of overgeneralisation when a word has a different behaviour from other words in its word class

We use as the main descriptive formalism that of *local grammars* (Gross 1993, Gross 1997) formulated as recursive transition networks (RTN). These we can compile into finite state automata for efficient processing. We can also integrate this formalism with *grammar patterns* (Hunston and Francis, 2000) as described by Mason and Hunston (2004) and Mason (2004). The latter will be the main focus of this section.

5.4.1 Related Work

5.4.1.1 Local Grammar

The term *local grammar* has so far been used in two different (but related) senses. Gross (1993) and (1997) uses it for a formalism (based on regular expressions) to describe the localised environment of a lexical item or a group of related items, especially intended to

deal with frozen expressions and synonyms. Furthermore it has the potential to apply *transformations* to capture related forms. Gross contrasts this with a *global grammar*, e.g. a transformational grammar, which deals with sentence structure on a more abstract level, concerned with the combinatorics of different word classes. One can combine a set of local grammars to describe a larger subset of a language, and construct those grammars from re-usable modules.

Sinclair and Hunston (2000) describe the second type of *local grammar*: instead of describing the syntactic environment of a word form it links up form and meaning, so that one would speak of a 'local grammar of evaluation' rather than a 'local grammar of *Bob lost his cool*-type sentences' (Gross, 1993). We can observe a certain overlap when Gross refers to a local grammar of date expressions, but it appears that Gross approaches the issue from the *form*-angle, whereas Sinclair and Hunston look at it from the point of view of *meaning*.

Gross's main point is how to represent local grammars in terms of finite state automata (FSA). Chomsky (1957, 21) explicitly ruled out this formalism when he stated that *English is not a finite state language*. However, Chomsky's objections are questionable for several reasons: first, focused only on *competence* he ignored practical restrictions related to *performance*, which in effect allow us to describe authentic language using finite state techniques; and second, since a local grammar describes localised phenomena only, non-finite elements such as long-distance dependencies or infinite embeddings do not actually pose any problems.

Sinclair and Hunston use a pattern approach to describe their local grammars, but there is no reason why we should not combine the two approaches, since we can easily express the kind of pattern employed in an FSA. FSAs can be used recursively, so they

suit this purpose well.

5.4.1.2 Pattern Grammar

Another related area concerns *pattern grammar* (Hunston and Francis, 2000), in fact quite close to a Gross-type grammar. Hunston and Francis express the typical syntactic behaviour of a word in the form of one or more patterns. Again, we can easily convert the pattern into an FSA. This has the additional advantage that we can combine all patterns of a word to form a single FSA, approaching the kind of description we find in Gross (1993).

Pattern grammar deals with sentence structure in a linear way. Rather than having a hierarchical structure of constituent parts as derived from immediate-constituent-analysis, it describes a sentence as a sequence of patterns, often 'flowing' into each other (through overlaps). This *pattern flow* is particularly suited for the description of spoken language, but works equally well with written texts.

Hunston and Francis remain sceptical about automating the processing of patterns. Their main objections (2000, 67), together with possible solutions are:

1. a computer program would not be able to distinguish between different uses of *that*, as in

    - *Daniel didn't miss the look of annoyance that flickered on Brenda Goldstein's face.*
    - *If anything, my mood is more one of annoyance that we haven't been winning when we have played so well in so many matches.*

In the first sentence, *that* introduces a relative clause, whereas in the second one it is an appositive clause. This should be easily resolved, simply because the relative clause starts with a verb, whereas the appositive clause has a noun group (here actually a pronoun) as its first element. A simple test of whether the following element is a finite clause should clarify the situation.

2. in some cases a *to*-infinitive is part of a pattern, whereas in others it is not:

   - *But then things started to go wrong.*
   - *A group of young children passing by stopped to watch us.*

   The second example here is actually ambiguous: the children could have stopped doing whatever they were doing at the moment in order to start watching us, or they could have been watching us while passing by but then stopped to watch and did something else (while still passing by). This requires more contextual information to be resolved.

3. the word *as* is ambiguous between a preposition and a conjunction:

   - *I went along dressed as a Japanese lady.*
   - *Rock queen Tina Turner didn't feel quite dressed as she stepped aboard Concorde yesterday.*

   Here the situation is similar to Hunston and Francis' first example in that it can be resolved by investigating the following structure: the prepositional phrase simply consists of a noun phrase following the preposition, whereas the conjunction is followed by a finite clause.

Taken together these examples demonstrate that the current capabilities of natural language processing can easily be underestimated. Even some simple techniques (such

as checking for the presence of a finite verb form) can easily distinguish between the problem cases presented. While a parts-of-speech tagger may not pick up those distinctions straightaway (e.g. in the case of *as*), a basic post-processor will do.

Other problematic cases, such as identifying the word that a pattern belongs to are difficult only because Hunston and Francis (2000, 68–71) presume a *careless observer* and *too cursory a glance at the concordance lines*. In some cases it is necessary to check whether a *that*-clause actually belongs to a word preceding the word in question, so for example with *Rumours had been rife that if war came...*, *rife* does not have the pattern **ADJ that**, but instead it is *rumours* with the pattern **N that**. This type of sentence is easily analysed correctly when doing a comprehensive analysis. The pattern of RUMOUR would be found from other instances where it was not followed by *rife*; if there were too few of those to be significant one could probably posit a frozen expression *rumours/speculation/suspicion/concern <be> rife that*. One will often misinterpret a sentence in isolation, but one of the main strengths of a corpus-based approach is that repetitions of the phenomenon under investigation will enable one to filter out any 'random noise' introduced either by idiosyncratic usages or by 'merged' patterns which are hard to separate out. If the sequence *rife that* only ever occurs with words such as *rumour(s)*, *suspicion*, and *concern* (all with similar discourse prosodies!), then it would not make sense to ignore this, and the common pattern **N that** of these words would lead to the correct interpretation.

Another example is the confusion between **adj** *enough* **to-inf** and *it* **v-link ADJ to-inf**. In this case the principle of longest matching, where a long pattern takes precedence over a shorter one, would lead to the right result.

In conclusion, Hunston and Francis significantly underestimate a computer pro-

gram's ability to identify patterns correctly. Even without sophisticated algorithms most of the problems they present are not at all *beyond the capacity of current computer programmes* (2000, 71). They probably wrote this with the computer as a simplistic (lexicographic) search engine in mind, where a program counts or retrieves instances of a given word and pattern combination. It is a different matter if the computer is used to try to identify patterns automatically using techniques from computational linguistics.

5.4.2 Patterns and Local Grammar

The manual identification of grammar patterns for any given word is a time-consuming and labour-intensive task, just like the identification of discourse prosodies described earlier. It is quite hard to automate this task, as Hunston and Francis (2000, 71) state *that frequent co-occurences of words do not necessarily indicate the presence of a pattern*, which requires interpretation of concordance lines by a human analyst. The best a computer will be able to do is a description based on frequency, as there are no other criteria by which to tell whether some sequence of elements is actually a pattern or not. Alternatively, rather than 'creating' patterns from nothing, the computer could be provided with a list of known patterns as a kind of 'seed'; if the recognition of patterns works reliably then the identification of a word's patterns (including their respective frequency counts) should be a doable task.

Given the possibility that sometimes more than one pattern matches (Mason and Hunston 2004, Mason 2004), preference will be given to the selection of the 'correct' pattern, based on properties such as length and number of actual words as opposed to word categories (i.e. the more specific pattern is chosen over a more general one).

5.4.3 Parsing Strategies

Unlike Brent (1993) we will process the input text with a parts-of-speech tagger, so that word class information is available for easier processing. Even though that introduces an element of error, the advantages outweigh the risks of wrong decisions. In the section on evaluation we will revisit this question and assess how far tagging errors influence the overall result.

With part-of-speech information available, there are a number of possible computational methods for discovering syntax patterns in running text. Initially, a chart parser would be used to recognise possible phrases and clause candidates. These are added as links in a chart. We can then encode grammar patterns in a finite state automaton which can then be used to recognise patterns in running text. The formalism for representing these FSAs could be the one described in Gross (1993). Alternatively we could use other, similar kinds of pattern recognition algorithms.

There are a few problems with this approach for identifying an unknown set of patterns:

1. we do not know initially what constitutes a pattern, so we do not know where the pattern boundaries will be. A possible solution would be to process grammar in the same way as described above for chains: instead of counting n-grams of word forms we would use n-grams of part-of-speech tags or phrase labels, and we would assume that fixed patterns rise to the top of the list on the basis of their frequency. That would approach Sinclair (1996b)'s second lexical relation, *colligation* (see section 3.1.1 on page 77). But we would also have to take into

account Hunston and Francis (2000)'s caveat that frequency of occurrence does not necessarily mean that we are looking at a real grammar pattern.

2. even if we have a list of possible pattern templates (e.g. the list of patterns in Francis *et al.* (1996) or the entries from Sinclair (2001)) we will find that more than one pattern would match a given situation. Any verb would have the potential pattern **V**. Every time we can match the pattern **V n** *to* **n** we also match the pattern **V n**. Mason (2004) describes some heuristics which were used to identify the correct pattern in a recognition task, such as longest-match, or 'lexical item before abstract category'. But unlike the recognition task described there, in this case we do not have the complete set of possible patterns of a verb, and will thus face increased uncertainty with multiple matches.

5.4.3.1 Chart Parsing

As we do not know what kind of elements (lexical items, word classes, phrases) are most appropriate for the task, we need to keep track of a number of potential candidates. This is especially important when we use clauses as potential complements: we cannot be certain that we have identified a clause with only a shallow approach to syntactic analysis, so we have to use cues that indicate the possible presence of a clause (such as the sequence 'NP VP (NP)' following a VP or the word *that*). At the same time we do not want to discard the first NP which—instead of being the hypothesised clause's subject—might be the 'real' complement/object in question, so we need to be able to store parallel choices. The most appropriate data structure for this purpose is a *chart*.

A chart is a table listing possible interpretations of items in a sequence. Charts are often implemented as networks/graphs, since they are more flexible. Winograd (1983,

119) explains that a chart describes *a record of all constituents and partial constituents produced in the course of recognition*. What we are using in the system described here is actually an *active chart*, in which pending (i.e. partially recognised) constituents are entered as well as completed ones. The chart contains a number of vertices, which represent the spaces between the words of the input sentence. Vertices are used to specify the range of a constituent. The other element of the chart is a set of edges, which connect two vertices and have a label, the name of the constituent they represent.

A top-down chart parser tries to insert prospective constituents as active edges, which contain a record of their own constituent elements which have not yet been recognised. If they can be found in the chart, the edge becomes a completed one, otherwise it is rejected, as the constituent could not be found. In traditional syntactic processing the aim would be to find a constituent 'S' which covers the whole sentence; in our case, however, we are looking for either a pattern sequence (in the case of pattern identification) or simply a range of phrasal components (in the case of colligation identification, see section 5.2).

Following the two approaches outlined above we create a chart in the first processing step using a traditional context-free grammar represented as a recursive transition network (RTN). That is adequate for this type of analysis, as it is only the higher levels where sentence complexity interferes with the performance of a CFG parser. In a second processing step we can then go through the chart and look for possible pattern matches, using the existing inventory of patterns as templates.

For the chart parser we can use a simple grammar, similar to the one used by the chunker for identifying usage patterns (figures 5.2 and 5.3). The chart parser uses a number of automata/RTNs for the different elements of patterns; these elements are

noun phrases, verb groups, to-infinitives, clauses, wh-clauses, quotes, and amounts. It is no problem that these elements can overlap, as they are stored in a chart which can deal with overlapping arcs. The pattern recogniser simply selects those arcs which are required to match a pattern.

This grammar is far from complete when it comes to describing the structure of a sentence, but again we are not interested so much in a complete structure, but rather in the identification of constituent phrases against which to match grammar patterns. A similar approach was followed by Niedermair (1986) in an automatic speech recognition system (SPICOS), as parsing spoken language is notoriously difficult. The parser described there also has a first step where nominal and verbal phrases are identified, and only in a subsequent step is an attempt made to combine them into a complete analysis of the utterance. At that stage higher level information about case frames can be taken into account, which is difficult to integrate in the earlier parsing step.

5.4.3.2 Pattern Recognition and Identification

In this first step we assume that we have a list of potential patterns available. As a source we are using a machine-readable version of the Cobuild dictionary (Sinclair, 2001).

The main difficulty with recognising patterns automatically is that the patterns are not restrictive enough, which means they can match a word sequence accidentally. This is more of a problem for nominal patterns, where any postmodifying prepositional phrase could be wrongly identified as a pattern. For example, the noun *decision* (Sinclair, 2001, 391) has the patterns **N to-inf** and **N** *on* **n/wh**. As a result, any noun that happens to be followed by an infinitive complement or a postmodifying prepositional phrase introduced by *on* will be recognised as having one of those patterns.

The reason for this over-generation is that there is less variability (and therefore also less complexity) in nominal patterns, whereas verbal patterns are much more varied. That might be due to the different types of nouns, 'simple' nouns such as *scooter* and 'predicative' nouns (so-called by Gross (1982) following Zellig Harris' terminology) which take complements like verbs do. In verbal patterns with different complement options it is much more difficult to match a pattern accidentally, apart from the shorter ones such as **V n**. So restrictions in co-selection do not extend only to lexical choices, but also to grammatical environments, an observation also made by Householder (1982).

If it is known which patterns a word has, then accidental identification is not a problem. In an (unpublished) pilot study the phrase *decided on [noun]* was manually investigated, as an anonymous project reviewer had claimed that one could not disambiguate between *They decided on the boat* (location) and *They decided on the boat* (rather than taking the plane). In practice such ambiguities are avoided by human speakers, presumably due to the risk of misunderstanding, i.e. if there is a pattern, structurally identical non-pattern uses are 'blocked'. Prepositional phrases involving obvious adjuncts (*they decided on Friday*) are not problematic: here a very small closed set of words leaves little room for doubt. This is another argument against intuition-based analysis, as many things are possible but are not used in practice. But without recourse to real data we would not be able to discover this and would postulate a difficult disambiguation task that does not actually exist.

However, as we want to identify patterns without knowing the set of correct answers in advance we have a problem. We can only hope that 'accidental' patterns will not occur often enough to overshadow the 'real' patterns, and that a simple frequency filter will remove the undesired ones. In order to assess the quality of the pattern identification we will evaluate it against the Cobuild dictionary, which provides patterns for its

entries. One difficulty here could be that the Cobuild dictionary does not list patterns exhaustively, choosing only the more frequent ones. This highlights the general issue of defining what patterns a word has, as Hunston and Francis (2000) do not specify objective criteria for that.

The main difference between the pattern identification and the previously described work on colligation is that colligation includes a general mixture of grammatical and lexical categories. Grammar patterns have a more limited vocabulary, which consists of grammatical categories plus a few common prepositions. These are used instead of the more general label `prep` when the choice is restricted to one particular preposition. Finding patterns is a more difficult task than simply *recognising* them, as we are less restricted in the set of candidates for potential patterns. While the pattern recognition task had a specified search space, namely the set of known patterns, the search space of the pattern finding task is essentially unlimited, apart from constraints on pattern length.

Unlike earlier studies (Mason and Hunston 2004, Mason 2004) we are not interested in the actual extent of the pattern in the text, but only in the result of the matching process. The pattern matcher therefore returns a set of patterns that could be found in the chart.

5.4.4 Evaluating Grammar Patterns

Altogether we have now distinguished two separate tasks:

1. recognise patterns in the environment of words, assuming that the set of grammar

patterns of a word is limited and already known,

2. identify patterns in the environment of a word where no restrictions are put on the potential patterns that can occur (but the total pattern inventory is limited and known).

We can evaluate fairly easily the success of the first task by comparing it to existing pattern descriptions such as Francis *et al.* (1996) and Sinclair (2001), as well as by inspecting the results manually in case the patterns were too infrequent to be included in those sources. Evaluation is easy because we know what to expect and we know how words and grammar patterns relate.

With the second task evaluation is not straightforward. We are again confronted with the 'discovery dilemma' of not being able to integrate our findings with traditional frameworks. Unless, that is, there is a good match between traditional categories and the empirically discovered ones.

In this section we will investigate a number of words and compare the patterns *identified* for them with the ones listed in the Cobuild dictionary. It is not as unproblematic as it seems: a genuine pattern might not be listed in the dictionary, due to a frequency bias in the corpus. We will use the written component of the BNC in order to approximate the setup used for the creation of the pattern grammar, namely a large general corpus.

As a starting point we will use the complete list of patterns (of all words) as listed in the dictionary. This is to reduce the overall search space; and we will further have to assume that this pattern list that we start with is complete, i.e. that it contains all possible patterns.

There are two major potential problems with regards to the evaluation of the approach:

1. Sometimes patterns are used in a 'non-canonical form', where the usual word order is changed for stylistic reasons. The use of the passive voice also makes it difficult to identify patterns correctly. Here we would suffer from reduced recall.
2. Patterns that are not in the dictionary can be found for two reasons: either they have been left out of the dictionary, or they have been matched accidentally.

We will have to bear the problems in mind when looking at the outcome of the procedure.

The first word we will be looking at is DECIDE. The frequencies of the inflected forms in the written part of the BNC are shown in table 5.8:

decide	5,815
decides	816
decided	14,201
deciding	1,841
Σ	22,673

Table 5.8: Frequencies of DECIDE in the written part of the BNC

The first point of interest is the frequency bias towards the past tense form, which makes up more than 50% of the total.

5.4.4.1 Patterns in the Dictionary

The Cobuild dictionary gives the following patterns for DECIDE:

- V to-inf
- V that
- V *against/in favour of* n/-ing
- V wh
- V
- V n
- V-ing
- V n to-inf

These patterns are distributed across five senses of the word. Since the senses are arranged in frequency order, the list of patterns found in the corpus should also be roughly in the same order of frequency of occurrence.

5.4.4.2 Patterns in the Corpus

We will apply the ten-percent filter for the following list of patterns identified in the corpus, i.e. each pattern that occurs with a frequency of less than ten percent of that of the most frequent one will be discarded. We will, however, list patterns from the dictionary list regardless of frequency, to give an indication of where they were in the list. In table 5.9 we have listed in separate columns the patterns identified for all the inflected forms of DECIDE.

	decide		*decides*		*decided*		*deciding*
2290	V wh	329	V to-inf	6724	V to-inf	214	V-ing
1048	V to-inf	119	V that	2781	V that	60	V n
847	V prep	119	V wh	2225	V *that*	51	V wh
628	V n	117	V n	1830	V prep	50	V to-inf
421	V *on*	79	V *that*	1772	V n	29	V prep
385	V that	76	V prep	1188	V pron	25	V *on*
286	V *how*	50	V pron	790	V-ed pron	22	V *on* n
285	V *that*	42	V cl	...	...	...	...
...	...	...	...	232	V wh	16	V that
14	V *against* n	6	V *against* n	144	V *against* n		
3	V *against* -ing	2	V *against* -ing	66	V *against* -ing		

Table 5.9: Grammar patterns for the inflected forms of DECIDE

Overall the result is very encouraging. The most important patterns from the dictionary have clearly been identified, even though their distribution varies considerably across the inflected forms. The pattern **V** *against* **n/-ing** is not frequent enough to make it through the filter.

There are some spurious patterns, and some duplication: **V that** clearly includes **V** *that* , the difference being that in the former pattern the *that* can be omitted.

We have attempted to determine the patterns of a word by matching from the complete inventory of all possible patterns those that occur with the word in question. From the result of the case study we can see that this approach provides a good way of identifying the patterns which can be associated with a word. That seems to contradict the predictions of Hunston and Francis (2000), who are pessimistic about the possibilities of doing so by computer. However, it leads to the main issue with grammar patterns, the lack of an objective means of determining what the actual patterns of a word are. In the absence of such a criterion we can only conclude that words do not have a fixed set of patterns associated with them, but rather tend to occur with a certain range of patterns.

There are a number of caveats that need to be taken into account when looking for patterns: one of them is that we might match a construction that on the surface looks like a pattern, but is not one in reality. That throws up the question of what we are actually dealing with: is it a description of (observable) surface structures, or of (unobservable) deep structures which are hypothesised by grammarians in an attempt to explain difficult examples? In section 2.3.2 we mentioned the example of *eager to please* and *easy to please*, which differs by only a few letters on the surface, but has a completely different underlying structure. The structure is linked to the different behaviour of *eager* as opposed to *easy*; and other words which share their respective behaviour also exhibit the same structures, for example *keen* or *willing* for *eager* and *hard* for *easy*. It is thus not surprising that the assigned structures differ considerably.

On the whole it seems unlikely that there are many cases where an identical surface structure leads to very different interpretations, as it would put too much cognitive load on the 'decoder' of the utterance. That does not mean that such cases will not exist, and in fact they are common in puns or jokes where initially a certain structure is suggested which then turns out to be a mis-interpretation. But it is not a basis for efficient communication.

This issue ties in with the description of spurious patterns by Hunston and Francis (2000), as described above (5.4.1.2) with the example of *rumours were rife that*. If this structure is repeated frequently, why not treat it as a fixed expression with its own patterns? The phrase will have a complementary distribution to other variants without the *were rife*, so that it can easily be treated as a different case. In the end it is unsatisfactory to invoke special circumstances in order to make up for situations where the basic description of a phenomenon falls down. It would be better to accept some degree of proliferation with pattern numbers in exchange for full objectivity in the application of

the pattern formalism to authentic data.

5.4.4.3 Finite State Patterns

Just like a local grammar as introduced by Gross, a set of grammar patterns can be expressed as a set of nodes, which are interconnected by arcs (Gross, 1997) to form a network with exactly one entry point, and exactly one exit point. Each node can itself have any number of entry and exit points. A node matches a certain type of inputs, which can be lexical items, lemmas, word class categories, or even a link to another network. The node basically contains a list of items it can match, which can be any combination of those items. An example of such a network is figure 4.7 on page 166; there it was used to represent multi-word units.

Such a network can be implemented as a list of nodes, where each node contains links to its possible successor nodes, while a grammar is a set of networks. Each network has a unique identifier which can be used to refer to it from within other networks. Matching a grammar is then a matter of selecting a network, and identifying a path through it from the entry to the exit points through nodes which match the appropriate input elements. INTEX (Silberztein, 1993) is an interactive system to create local grammars, which are stored in a plain ASCII format; these local grammars can then be applied to corpus data.

Finite state automata as a means of representing grammatical information would be a useful tool for the overall analysis of the linguistic description attempted in this project. Several components, such as multi-word units or grammar patterns, can be expressed in the form of FSAs. One of the original aims was to generate output suitable for processing with INTEX, but because of lack of time it had to be moved into section

7.4 on future work.

5.5 Grammar: Summary and Evaluation

In this chapter we have looked at three aspects of empirical grammar: colligation, usage patterns, and grammar patterns. While colligation is a vague concept that still needs a lot of further research and elaboration (like the closely related area of collocation), it can still be useful for exploring the interface between lexis and grammar. It can make accessible subtle tendencies in the grammatical environment which can easily be missed with a purely rule-based approach which focuses on phrase structure.

Usage patterns have problems with their overall coverage of syntactic relations: owing to their shallow nature, they miss more complex structures and are vulnerable to mistakes in the analysis. Nevertheless, given the minimal effort they require, they can still provide useful information about typical collocations within syntactic relations.

Grammar patterns, finally, are the most sophisticated of the three aspects presented in this chapter. We have investigated ways of recognising patterns automatically, given an existing inventory of patterns from a dictionary. Despite predictions to the contrary, recognition works well, and can now be performed on a larger scale for comparative studies. Grammar patterns can also be used to analyse further the correlation between form and meaning: we can assume that different grammar patterns are used with different senses of a word, and by identifying the pattern we can try to verify this by comparing the semantic features of instances of the various possible patterns.

In conclusion, we are now able to describe various aspects of a word's grammatical

behaviour in a fully automated way.

CHAPTER 6

MEANING

6.1 Introduction

The study of meaning is a permanent interest of scholarship. *J.R. Firth*

Most work in corpus linguistics has so far avoided semantic issues, despite Firth's statements about lexical semantics going back half a century. Researchers instead concentrated on areas which initially benefitted from the ability to access corpus data, such as lexical studies (e.g. Sinclair *et al.* (2004), originally published in 1970) and, later, phraseology. References to the meaning of words are generally limited to general 'prosodical' aspects, such as positive/negative in Louw (1993).

As already mentioned in section 2.3.4, there are doubts that meaning is ever accessible through empirical methods. Arguably the methodology for investigating meaning from corpora is not advanced enough, and the meaning of a word has been only vaguely described through the set of its collocations, as for example in Stubbs (2001). This does

provide a good indication of a word's uses, but ironically suffers from a problem identified only through collocational analysis itself: the word (as in 'orthographical unit') is not the primary unit of meaning. Thus, describing the meaning of single words is problematic in the first place, and describing it using other single words is less than perfect.

In a critique of formal or 'scientific' approaches to the study of meaning Sampson (2001, 206) states that *word meanings are not among the phenomena which can be covered by empirical, predictive scientific theories*. He bases his argument on the observation that meanings are internal to a speaker and constantly change in unpredictable ways. However, according to Stubbs (2001, 35) *[t]he vocabulary of a language is not an unstructured list of words* but rather *is internally structured by many clusters of words, which stand in different relations to each other*.

So while Sampson may be right that we cannot predict changes in word meanings, he is clearly wrong in denying that we can use empirical methods for the study of meaning: if a language's vocabulary is 'internally structured' then this structure must be reflected in some way in language use, a situation somewhat reminiscent of the cave allegory in Plato's *Republic*. Here the shadows on the wall reflect what is happening 'in reality', i.e. they correlate with the movements of whatever is causing them.

And while Sampson is certainly right in saying that meanings are internal to the speakers of a language, language, being a representation of thoughts and concepts, is not completely unrelated to the outside world. We can introduce a duality of the outer, 'real' world and the entities contained in it, and its representation in the mind of a speaker, reflected through language (among other media). It is more than conceivable that the structure of the outside world will in some form be reflected in the network of

word meanings, at least where words refer to external entities. Analysing the relationships between words we ought to be able to identify systematic relations which must be isomorphic to 'reality'. The link between the 'inner' and 'outer' worlds in itself is (currently) out of the reach of linguistic study.

But arguably, from an empirical point of view, the outer world exists only in the individual's perception, or at least can only be talked about using language which is individual to a speaker. Since the main function of language is to transmit meaning through various channels (see Firth (1951)), meaning must be encoded within language somehow, and the elements of it which carry meaning can be investigated. What we will not be able to determine is the *interpretation* of those elements, which happens inside the mind.

In this chapter we are looking at meaning and how we can extract aspects of it from corpus data. For the empirical analysis we will have to restrict ourselves to lexical semantics, and we will attempt to describe the meaning of words by investigating the internal structure of the vocabulary. The underlying principle is the correspondence between form and meaning, which has been noted in previous chapters.

6.1.1 The Method

We will explore three ways of describing the structure of the vocabulary through shared environments. The environment will first be defined through collocation, and similarity of collocations ought to be an indicator of similarity in meaning. Second, we will look again at usage patterns. This time we argue that words which share 'partner' words in the same syntactic relations also share aspects of meaning. As we said in the last

chapter, usage patterns can be viewed as an extension of collocation, so this is a natural progression. And finally, we are going to look at directly shared contexts. We will extract the contexts of a word through multi-word units (see section 4.3), and will identify words which can replace the node word in the same MWU. The more shared MWUs two words have, the more the words will have in common, as they occur in a larger number of common contexts.

6.1.2 Problems

So the common theme of this chapter is the analysis of shared contexts. The basic flaw in lexical semantics (as touched on above) is that we still use the (single) word as the basic unit of meaning. The same point has been identified by Grefenstette (1994, 144), who states that *[d]uring the course of this work, we have become convinced that restricting our work to individual words, [...] neglects a large portion of domain-dependent concepts that are expressed as multi-word terms.* Up to a point this limits the validity of our three approaches, as units of meaning are presumed to be above word level, and what we are actually looking at are component parts of units of meaning which may belong to a number of different units. However, in the absence of a workable definition of units of meaning we will have to compromise.

Also, even when looking at individual words which are elements of larger units we should be able to make some useful discoveries about aspects of meaning. A closer (manual) inspection might then lead us to larger units. Furthermore, some units of meaning might 'coincide' with single words, which we should be able to identify. However, evaluation of the outcomes is rather problematic, and will create more questions than answers.

In principle there would be no problem in applying the first method (collocational environment) to units of meaning, as we would simply collect the collocations from a set of concordances of those units. But a practical problem arises: few multi-word units have a sufficiently high number of occurrences to calculate enough meaningful collocates. If all multi-word units have only one or two collocates, then it will not be feasible to calculate the overlap between sets of collocates.

The second method (usage patterns) is not compatible with larger units of meaning, as it presupposes a traditional syntactic model of phrases and head words of phrases, which conflicts with the multi-word unit approach to higher level grammatical units. Multi-word units are not part of any definable syntactical relationships that could be expressed in usage patterns. Nevertheless, the method was included in this project in order to see how far we can get with it.

The third method (multi-word units) is based on units beyond word-level, but then inspects how elements of those units can be exchanged. The resulting semantically similar words are arguably related, in that they can be combined with the same context to form a larger unit. That would indicate a degree of freedom in multi-word units, in that individual elements can be substituted for others, with little or no change in meaning. Here we are still looking at individual word tokens (which occur in the same context), rather than the relationships between larger units. But at least we can separate out different senses of the word forms involved, provided they have a complementary distribution across the multi-word units.

This last method is supported by Firth (1952, 23), who states that *[t]he possibility of substitutions not amounting to commutation is an indication of similarity of value or function.* Firth is referring here to the substitution of words in a collocation, which in

his usage is a pair of words; the multi-word unit substitution thus goes one step further by taking into account a larger context than just one other word.

Steiner (2004) describes previous work on word class induction based on distributional approaches, which also makes use of contextual information to group together words with similar environments. The commonality between those approaches is that they postulate similar properties of words (here: syntactic behaviour/word class) on the grounds of shared environments. Steiner (2004, 60) states that *[d]a Kotexte auch von semantischen Faktoren bestimmt werden, sind die Grenzen zwischen syntaktischen und semantischen Klassen fließend* ("the borderline between syntactic and semantic classes is fuzzy, as co-texts are also influenced by semantic factors").

The fact that the context can be exploited to determine both semantic and syntactic classes somewhat reduces the validity of separating syntax and semantics with regard to a word's distributional behaviour. One could argue that word classes are both semantic and syntactic, depending on the granularity or level of detail of the classification. The more classes there are, the more important are semantic aspects in the classification. With a mere eight classes (as used in traditional grammar), only syntactic regularities have an influence. Contexts here are restricted to *function* words, and only a few templates (see Fries 1952) apply, which effectively define the shared contexts of the words to be classified. Once the contexts contain *lexical* words, and thus increase in number, fewer and fewer words will share the same contexts, and semantic aspects start affecting the grouping of words. Of course, lexical words and function words themselves are on a continuous scale (which roughly correlates with their frequency of occurrence), with no obvious boundary.

In conclusion, word classes can be seen as a continuum of groupings with any num-

ber of groups between one (meaning 'word') and the total number of tokens (with each token in its own class). Clearly, these extremes are of no use for a linguistic description, but this view of word classes provides a useful unifying view of a word's distributional behaviour, both syntactically and semantically.

Steiner (2004) also addresses the problem of *homonymy*, when only word *types* are clustered according to their distributions, rather than word *tokens*. There are obvious problems with this, especially in a language such as English, which has a large number of homographs: many word forms can represent a number of word classes, for example *take* can be a noun or a verb, and *light* can be a noun, a verb, or an adjective. Incidentally, capitalisation also matters here, since *March* and *march* are two entirely different words. We will address that issue when discussing the separate methods below.

6.1.3 Case Studies

We cannot easily compare the three methods, because they are designed to achieve different things: the first method evaluates a specified set of words to clarify the semantic relations between them, whereas the latter two try to find similar words from the whole corpus. In order to compare and evaluate those two 'open' methods, we will run a number of case studies. The words used for evaluation have been selected semi-randomly with a view to testing several aspects of lexical semantics:

dry This adjective has been chosen because it is an example from Stubbs (2001); it has an 'obvious' antonym in *wet*, but is frequently used in ways where it can be contrasted with other words, such as *sweet*. Stubbs uses it as an example of the need for extended units of meaning, as there is no single word antonym that

applies in all contexts.

plant This word is ambiguous in multiple ways: it can both be a noun and a verb, and it also has multiple distinct senses. The verb can, for example, refer to the planting of greenery, but also to the placing of surveillance devices or fake evidence to incriminate a suspect. The noun can refer to vegetation, heavy machinery and factories, as well as a fake member of the audience designed to help a conjurer or comedian.

brown Colour adjectives are expected to behave in a different way from other adjectives, so we would expect other colours to be chosen as similar. It would be interesting to see if both methods identify the same set of other colours as similar.

computer This is a fairly 'boring' noun, in that it is not ambiguous in contemporary usage.

Monday Similar to colour adjectives, days of the week are a closed set, and we would expect to find that the other members of the set are identified. It would be interesting to see whether any other words are similar in meaning.

Germany Proper nouns usually behave differently from ordinary nouns. With countries there is the additional aspect that they are often used as a metonym for their governments. So there should really be two kinds of similar words, other countries and words referring to administrative bodies.

Africa This is similar to the previous one, except that continents are an even smaller closed set, and there is not usually a matching administrative body covering the whole, apart from Australia.

Smith This is a plain proper noun; we would expect other names or personal pronouns to be identified as similar.

For these case studies we will be using a mixed corpus, trying to avoid a bias that might occur if a homogeneous corpus was used. This may add some statistical 'noise' to the calculations, and the results may be less clear-cut than they would be with a more restricted corpus, but it will also cover a broader range of possible usages. This is the set of corpora that is used for the case studies:

- BBC
- LOB
- Flob
- Frown
- MCon
- BLT

There are elements of news coverage, academic and general written language. The total size of the corpus is around 22 million tokens.

6.2 Collocational Overlap

The idea behind the collocational overlap method is that collocations are the words which occur within the context of a node word more significantly than expected by chance. Therefore the collocates of a word describe its context in a condensed way, and through analysing the distribution of collocates across the environment of a number of words we can assess their semantic similarity.

The starting point here is a set of words to be investigated. As the statistical method used for the analysis works best with a limited number of words, the candidates should be selected from ones that are already hypothesised to be similar. Ideally, we would scan the complete set of words and their respective collocates, and select as initial candidates those words which overlap in their sets of collocates. However, that presupposes a complete analysis of the collocations of all words; for technical reasons it has not been possible for this project, as the calculation of the data is too time-consuming at present with the available computing resources.

6.2.1 Starting Point

Instead, an existing set of collocations has been used to provide a starting point for the exploration of this method of determining word similarities: the COBUILD Collocations database, which was available as a computer-readable file. While it is not clear what parameters were used to compute the lists of collocations (see section 4.2.2), the exact details should not matter here. What the data provides us with is a list of words and their most significant collocates, and by investigating the overlap between these words we can make a first approximation of word similarity. We can then feed the sets of words deemed similar into our own procedure, where collocations are calculated differently, and see whether the results make sense; the operational principle is the same, even if the implementation is different.

The collocational overlap was determined for all words from the Cobuild database, and those with a distance of less than 0.8 were assumed to be similar enough. This cut-off point was chosen after a brief inspection of some sample data: most words tend to have a distance greater than 0.8, and words which seem to be genuinely related usually

have around 0.7. Only a few words have distances less than 0.6, which indicates that in general the range is fairly restricted.

The metric used to calculate the distance between two words i and j from the Cobuild database was

$$d(i,j) = 1 - \frac{2 \times |C_i \cap C_j|}{|C_i| + |C_j|} \tag{1}$$

in other words, twice the number of shared collocates divided by the sum of the sizes of each word's collocates. This is subtracted from 1 in order to yield a distance value, with 1.0 being the maximum distance, and 0.0 being identity.

6.2.2 The Procedure

The method used here to analyse the semantic relationships between a set of words first gathers the collocates of each word and collects them in a set. Then for each candidate the co-occurrence frequencies of each word from the set are counted, even though they will mostly be the collocates of the other candidate words. The counts are entered into a contingency table which is then analysed using Correspondence Analysis.

The outcome is a two-dimensional arrangement of both candidates and collocates. The way to interpret this graph is that the candidate words have their positions in the coordinate system depending on their affinity to any of the collocates; the two data sets are usually superimposed, but with different scaling factors, so that they have to be considered in relation to the origin of the coordinate system. The method will become clearer in the following section with the presentation of the case studies.

As a result we can see which of the candidate words are more closely related to

each other than to other ones, and we can also identify which collocates are responsible for the arrangement in the two-dimensional space. While the initial interpretation is done visually by a human analyst, it can easily be automated, as it involves simply the comparison of coordinate values.

One important point to note is that the result will be very different depending on the corpus data used. The Cobuild collocations have been computed using the Bank of English corpus, and using a less broadly distributed corpus to investigate semantic arrangements would most definitively yield a different outcome. This, however, is no cause for concern, as meaning is not fixed in the same way as phraseology, for example. Meaning as expressed through language is a mirror of the speaker's view of reality, and thus by definition subjective and bound to an individual. Corpora as collections of utterances from a large number of different speakers will give rise to conflicting interpretations of reality, or just emphasis on different issues. For example, the BBC corpus will talk about *democracy* in a different way from that found a corpus of philosophical debates. Hence one would expect different words in the neighbourhood of *democracy* in two such different corpora.

6.2.3 Case Studies

The first example is *abolished*, which is listed with the candidate words of *abolish*, *abolition*, *expenditures*, *repeal*, *introducing*, and *democratic*. Merely by inspecting this list we can see that there are several different logical relations at work which cannot be identified through collocation alone, owing to the spatial/grammatical differences.

A correspondence analysis (BBC corpus) yields the view in figure 6.1. This can be

compressed (and somewhat simplified) to the following table 6.1, where each row represents a candidate word (using the above terminology) and each column a collocate. The table indicates which collocates are 'attracted' by the respective candidate words, and as such they give rise to differences in usage.

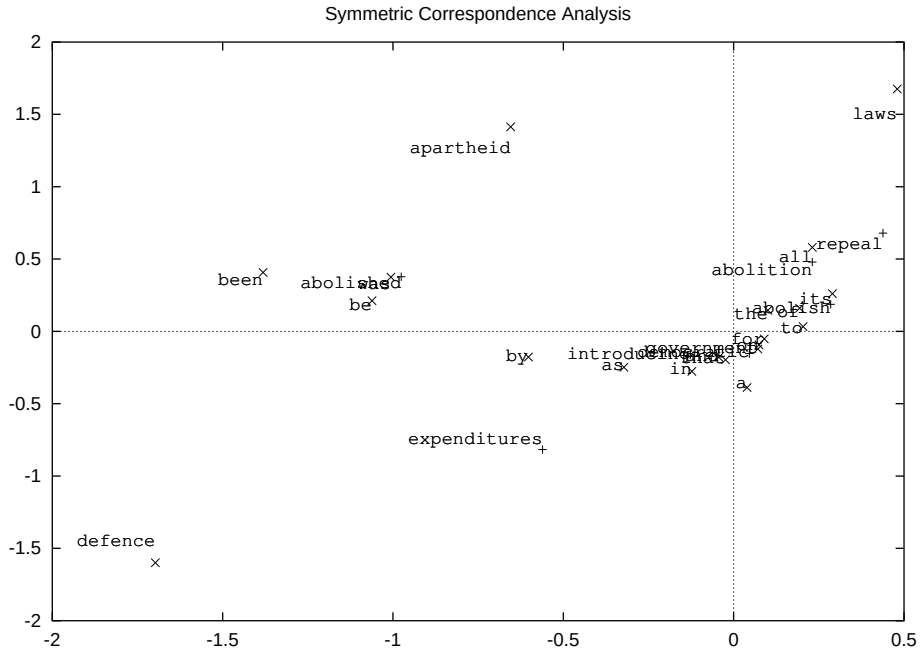

Figure 6.1: Correspondence Analysis of *abolished* and related words

	laws	apartheid	defence	by	all	its	government
abolish	X				X	X	
abolition	X				X	X	
abolished	X	X	X				
repeal	X	X			X		
expenditures			X	X			
introducing			(X)				
democratic							(X)

Table 6.1: Collocations of *abolish* and similar words

Here we can see the difference in usage/meaning between *abolished* on the one hand, and *abolish* and *abolition* on the other; the two final candidates, *introducing* and *democratic*, do not have strong tendencies towards any of the collocates, and in fact, *repeal* is the closest word in meaning to *abolish* of the set.

Using a different significance function (t-score instead of raw frequency) we get a slightly different picture: here, *expenditure* is drawn away from the other candidates through the collocates *cut*, *defence*, and *unexpected*; this is so extreme that the display of the remaining candidates is too compressed for the analyst to notice anything. Re-running the analysis without *expenditure* we get a similar picture to the one described above. One additional collocate, *visas*, is in a similar position to *law*, but the overall arrangement of candidate words in relation to each other remains unchanged.

Looking at another word, *ballet*, this time using the written part of the BNC as a corpus, we again get an outlier that distorts the overall display, *rhythms*, associated with the collocate *daily*. Without this outlier the picture in figure 6.2 shows a reasonable arrangement that can be interpreted fairly easily. Words that seem intuitively connected are displayed in proximity.

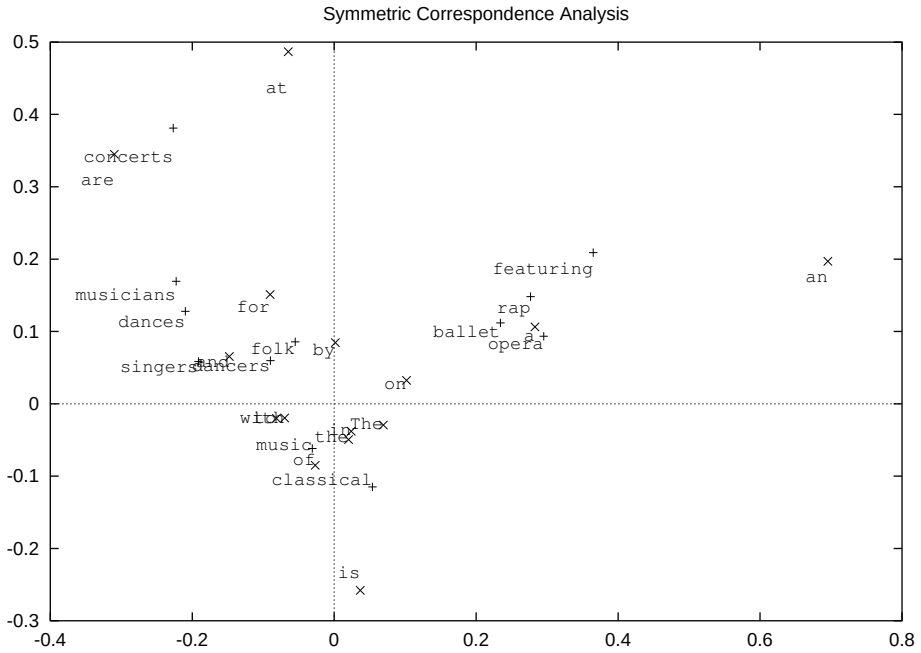

Figure 6.2: Correspondence Analysis of *ballet* and related words

A further example is *rights* (also using the written part of the BNC). Here we find an

outlier as well, *fundamentalist*, with the collocate *Islamic*. A second iteration identifies *umbrella* with the collocate *under*, a third *separatist* with *Basque*. The next step then gives *liberation* (*national*) and *liberties* (*civil*), before we reach the result shown in figure 6.3. Unfortunately most of the collocates still remaining are grammatical words.

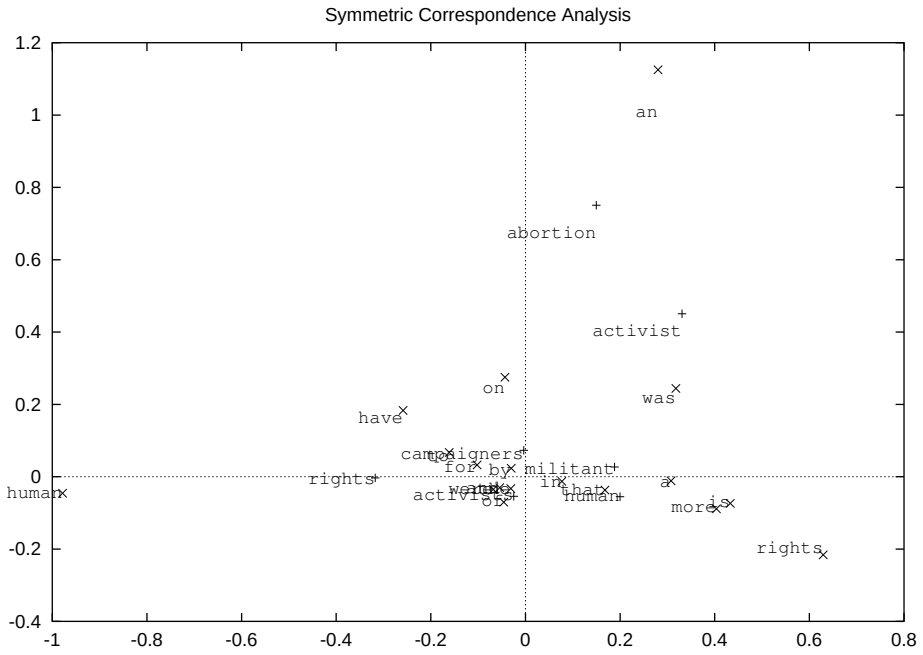

Figure 6.3: Correspondence Analysis of *rights* and related words

In general, many of the words extracted from the Cobuild database do not seem to have much in common. A more suitable application for the collocational overlap method could be to look at related forms, e.g. inflected variants or derived words. A partial example is *abolish* above; we will now briefly look at the full set of *abolish*, *abolishes*, *abolished*, *abolishing*, and *abolition*. Using t-score as the significance function results mainly in function words as collocates, and then we get the obvious attraction of *of* with the noun *abolition*, and past tense forms of *be* with *abolished* (see figure 6.4). Here we have a straightforward correlation between collocation and colligation, but it does not say anything about semantics. Resorting instead to mutual information, which tends to favour rare words and suppress grammatical words, we get three completely disparate

groups: *abolition*, *abolishing*, and the remaining *abolish*, *abolishes* and *abolished*. The problem is that the words are too specific and 'random' and there is no overlap at all.

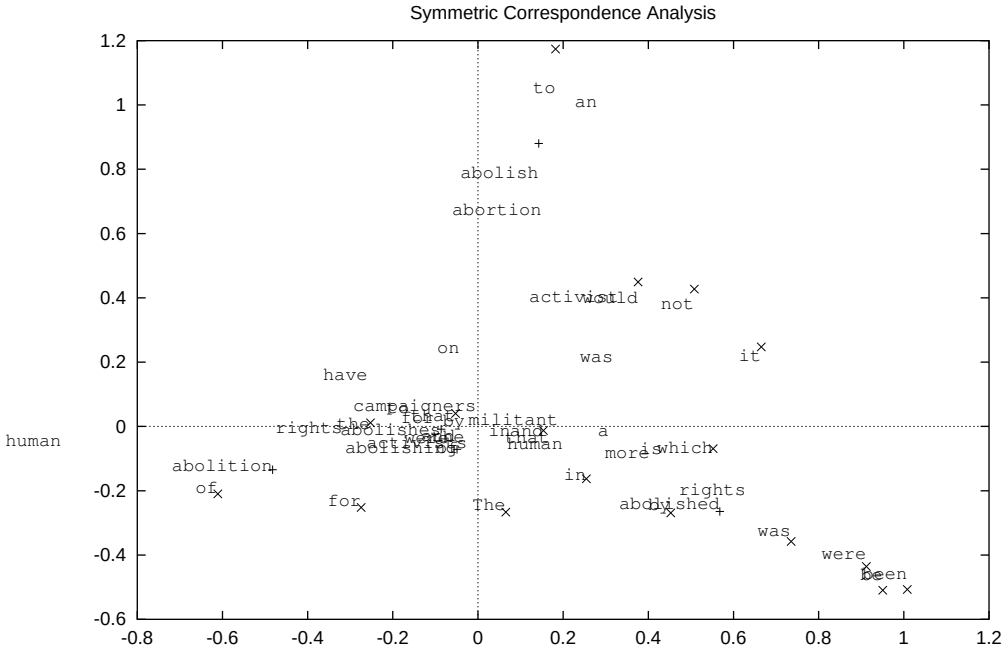

Figure 6.4: Correspondence Analysis of ABOLISH and related words

6.2.4 Evaluation

The collocational overlap method of displaying the semantic structure of a set of related words is useful, as it provides an explanation for the spatial arrangement of the words through the collocates superimposed on the two-dimensional graph. There are a number of limitations: the method cannot easily be used with a large number of elements, as the display becomes hard to read. However, this is not a problem if it is to be analysed automatically, as then only the numerical values are important, not the readability of labels that overlap. The procedure is also sensitive to outliers. But outliers can easily be recognised, and then the procedure can be repeated with the outlier removed, which gives a clearer picture. In general, the choice of words to analyse needs to be sensi-

ble; the approximation used here (the degree of common collocates from the Cobuild collocational database) is clearly not ideal.

Another problem is that the collocates often tend to be grammatical words, and then the identified structure simply mirrors syntactic regularities rather than semantic ones. That could probably be fine-tuned by using different parameters for calculating the collocations, and perhaps a frequency filter or stop word list. It is conceivable that different applications demand different ways of computing the words that define the environment of a target word: for a more syntactic view we need grammatical words, whereas semantic applications might work better with a bias towards low-frequency or content words. This issue is also an argument against a generalised 'one size fits all' approach to collocation.

6.3 Usage Patterns Revisited

6.3.1 Introduction

As described in the previous section, one problem with context described through collocates is that they are the result of a number of linguistic processes working in parallel. Two words might collocate with each other by pure accident, or because they belong to certain word classes or have certain phonological features (for example to form alliteration), not necessarily because they share aspects of meaning.

A more 'concentrated' relationship is captured in the usage patterns that were used to describe the grammatical behaviour of a word (see section 5.3). Here we are looking at word pairs whose adjacency is motivated by a syntactic relation, rather than

just through proximity. Hence they are more suitable for defining an environment than collocations.

In the previous chapter we saw how the other re-current elements in a usage pattern can be identified; in this section we are going one step further to compare words which have similar sets of 'partners' in one or more relationships. To keep matters simple we will refer to the 'other' word in a usage pattern as an *argument*, even if such words are not arguments in the traditional sense.

The approach is inspired by the work of Schwarz *et al.* (1991). In an information retrieval context they were interested in identifying words which were similar to a search term in order to expand a query. They used a shallow parser that identified head-modifier relations between words, similar in style to the usage patterns applied here, but on a larger scale and with more coverage. The similarity of two words was then evaluated by the overlap between modifiers; in other words, two words which could be modified by the same set of modifiers were seen as similar. The reasoning is that similar entities will have many properties in common (expressed in language through modifiers).

More recently, Lin (1998) has used dependency triples and clustering to derive word similarity information. His dependency triples are similar to usage patterns described here, in that they comprise two participating word forms and a relation between those forms. He places the emphasis on making a full syntactic analysis of the data and explores similarity measures and clustering, building on work by Grefenstette (1994), Hindle (1990), and Ruge (1992). Lin also describes a pruning algorithm to discard unwanted results. One important point about his work is that he requires words to occur at least 50 times in the corpus, so he either needs large amounts of text or can analyse

only reasonably common words. Thus it will not be possible to achieve a broad coverage of a corpus using his methods.

6.3.2 Procedure

From the procedure described previously (see 5.3), we have available a set of triples consisting of two (phrasal head) words and the label of the relationship between those two words. We also have, for each word, a set of significant arguments (the 'other' word in the relation). These arguments define the shared environment between the words under investigation, and they also provide a link to those words which themselves have one or more of these arguments. The latter set of words is the set of candidates considered as semantically related to our target word.

One relation is considered at a time; it is not immediately obvious how the results of analysing various different usage relations could be combined. From a theoretical point of view they represent different usages, and are quite likely to belong to different word senses. Taking a particular noun, it is conceivable that it shares the subject position in clauses (with a restricted set of verbs) with a particular group of other nouns. But shifted to object position, those same nouns do not necessarily also occur as object to a different set of verbs. A further step in the analysis could be to look at the similarity of the sets of similar words, and effectively assign a similarity value to the usage patterns of a word, which would give rise to statements such as: 'with (noun) X the SV and PN relations are similar'. This would mean that the semantic behaviour of word 'X' in those two grammatical relations would be similar, and could indicate a particular word sense of 'X'.

For each set of candidate words a set of arguments is calculated, and this set is then compared with the target word's arguments. (At present the process is repeated each time, but the arguments could be retrieved from a previously processed data set). The similarity between two words is then calculated through the overlap of the sets of arguments. In the output, the list of shared arguments is included to aid the understanding and interpretation of the results.

6.3.3 Problems

The problems with the usage pattern substitution procedure are mainly theoretical. It presupposes a grammatical framework, and so cannot be classed as fully empirical. Neither is it possible to obtain the grammatical relations which form the basis for the usage patterns in a fully automated way, since they rely on logical roles such as 'subject' and 'object' which have no empirical basis. It might be possible to substitute general surface relations, such as 'nominal chunk followed by verbal chunk'; these might work at least in fixed word order languages such as English, but would probably introduce too much statistical 'noise' into the analysis.

A further issue is that phrasal elements are reduced to their respective head words, so that *loading space* and *air space* would be treated the same. Again it is the single word *vs*. multi-word units problem that distorts the results. The phenomenon is often the cause of unexpected outcomes at this stage of the analysis.

6.3.4 Case Studies

Here we will present only a short summary of the eight case studies; the full XML output of the analytical procedures is provided in appendix C.

6.3.4.1 'Africa'

The first relation, **NN** with *Africa* as the first element brings up mainly dubious examples, which are most likely the results of processing errors. One problem with the word *Africa* is that it is not just a continent, but also part of a country name (*South Africa*), which causes interferences between the 'continent' usages and the 'country' ones. This becomes even clearer with the **SV** relation, where *being*, *France* and *party* are the top substitutes, based on verbs such as *expelled*, *granted*, and *denounced*. In compound nouns, the last noun element is taken as the head of the structure, which means that *South Africa* would be indistinguishable from the continent *Africa* on its own.

A further experiment was run with *Europe*, and here the results are surprisingly similar, with *countries* and *country* the most frequent substitutes. It seems that continents are predominantly refered to as political entities; however, we would expect different results if we had used a corpus of geography texts.

The **VO** relation works similarly, with candidates including *area*, *China*, *Baghdag* [sic], and *Britain*. Instead of other continents the majority of the substitution candidates are countries and capital cities.

6.3.4.2 'Germany'

The second case looks more promising: with **AN** we get *country*, *body*, *nation*, and *period*. The *body* usages go with adjectives such as *sovereign* and *neutral*, so they would refer to the 'political body'. The adjectives with *period* include *post-war* and *Nazi*, and are used to refer to different periods in recent German history.

The **NN** usage (initial slot) are rather successful, as they almost exclusively return other countries, administrative bodies, or politicians. Oddly there are also some adjectival substitutes like *German* and *Belgian*; these seem mostly to be caused by tagging errors, as they would not be parts of a nominal compound. On the other hand, they could be part of a larger structure that is also included in the **NN** relation, as for example in the (real) example *a West German of Arab origin*. Here *German* is the head of the first part, and the whole structure would fit the **NN** usage pattern.

The remaining two relations, **SV** and **VO**, have similar results.

6.3.4.3 'Monday'

NN (first position) gives all the days of the week, *week*, *night*, *Arabic*, and also *st*, *rd*, and *th*, clearly from being used in a date format. Later down the list some other, initially unrelated words, occur, for example *football*. Here the shared argument is *violence*, which presumably can be prefixed by *Monday* to refer to a particular incident. In the same vein we get *Jerusalem* with the arguments *killings*, *shootings*, *incident*, and *violence*. It would be interesting to explore whether the same applies to other days of the week; it seems obvious that not all days of the week are used in the same contexts, as weekly activities

are structured differently.

With **NN** in the second position we get just *Wednesday*.

6.3.4.4 'Smith'

NN returns a large number of surnames and initials, and one *reports*, which is presumably caused by the by-lines used in the BBC corpus, where reporters are named by their first name; the lack of context leading to *reports* being mis-tagged as a noun.

This outcome suggests and interesting application of this method for name recognition, where a list of common surnames could be used to seed a more general search for names not mentioned before.

6.3.4.5 'brown'

The **NN** relation yields a similar result to *Smith* above, as *Brown* can also be used as a proper noun. Again we see that even small details such as the upper/lower case distinction can be very important when looking at distributional behaviour.

Other relations are not frequent enough to produce results, which seems to be a serious problem with this method: 22 million tokens is a substantial amount of data, even if many current corpora are about an order of magnitude larger.

6.3.4.6 'computer'

With **NN** we get *you*, *space*, and *animals*, based on arguments such as *studies*, *model*, and *simulation*. Another set of substitutes is *car*, *French*, *Japanese* with arguments *maker* and *manufacturer*.

Other relations do not provide any further substitutes.

6.3.4.7 'dry'

There are no results for *dry* using this method.

6.3.4.8 'plant'

With **NN** the top substitutes are *plants*, *industry*, *supplies*, *company*, *factory*, and *failure*. They suggest that the 'industrial' sense of *plant* is the dominant one, though *failure* has as one argument *crop*, which points to the 'agricultural' sense.

VO (first slot) has *planted* as substitute, with the arguments *bomb*, *bombs*, *seeds*, and *trees*. This is an encouraging result, as the algorithm is not aware that the two word forms are morphologically related. It would now be interesting to see whether *bombs* are offered as substitutes for *seeds* on the grounds that they can both be planted. In the second slot (which is a noun) we get *aircraft* and *bank*, both of which co-occur with arguments *build* and *used*.

6.3.5 Evaluation

Considering that no semantic knowledge has been made available to the system, it does surprisingly well. A formal evaluation is problematic, of course, as the procedure is essentially explorative, and we do not try to re-create any existing semantic hierarchies such as WordNet (Miller *et al.*, 1990), which is often used for evaluation (e.g. in Lin 1998).

Clearly, looking at words in isolation is less useful than comparing a set of similar words. For a more complete evaluation we would need not simply a few case studies, but a comprehensive analysis. It was not possible to do that at this stage of the project because of the demands on computational power, but it will be a further step at a later point in developing the methodology.

So far the results are useful, as they deliver insights into the structure of language, but they cannot be accepted without careful consideration: clearly *computer* and *animals* are not synonyms, even though they share a number of syntactic environments. It is hard to say whether more data would yield better results, or whether it would simply introduce more dubious cases. Using a more homogeneous corpus might be one solution, as it would reduce the number of different contexts that influence the substitution procedure.

In summary, without further adjustments the output of this procedure cannot be used directly, but it can instead provide useful ways of studying certain (unexpected) word usages. More research is required to determine whether there are fundamental problems with the method, or whether it can be improved. Given the theoretical issues

mentioned above, it seems unlikely that a substantial improvement can be achieved.

6.4 Paradigmatic Substitution

6.4.1 Introduction

The usage pattern approach described in the previous section yields interesting and mostly plausible if sometimes unexpected results. But it presupposes a traditional view of syntax, namely one based on phrase structure, with a set of fixed roles within a clause (subject, predicate, object). Problems with the extraction of usage patterns were mentioned in the previous chapter (5.3). From a purist's point of view a more satisfactory means of analysing shared contexts would be through multi-word units as discussed in section 4.3. Here we also identify the context in which a word form occurs, but we do not make any assumptions about the relationship between the word form and the elements of that context.

6.4.2 Procedure

In a previous processing step we have already retrieved the multi-word units which are required for the paradigmatic substitution. Frames and chains are processed separately, since the different means of retrieving them could otherwise make interpretation of the results more difficult; we thus end up with two sets of multi-word substitution lists. Furthermore, only 'bounded' templates are used, which are those where the target word has at least one other item either side of it. Otherwise the substitution could match the beginning or end of a larger unit, which would distort the results.

For each set of multi-word units we then collect concordance lines. It is rather difficult to search for a phrase *without* a particular word, especially when the retrieval mechanism is focused on words as the search terms. So we compare the frequencies of the other words that make up the MWU besides the target word, and sort them according to increasing frequency. Next, concordance lines for the lowest frequency word are retrieved, and subsequently filtered to exclude those that do not contain all of the remaining words. Then, a wild-card pattern match is applied to the concordance lines to make sure that only the exact MWU template (with a blank instead of the target word) matches. All words that fit into the blank slot are collected in a (frequency) list. Single occurrences are discarded at this stage.

We now have a set of frequency lists, one for each MWU of the target word. Next we go through each list and inspect the candidate substitutes. Initially, all those that occur in the given template more frequently than the target word are discarded on the rationale that the target word would be a substitute for *them*, rather than them being a substitute for the target word. The same argument is also used for the distinction between upwards and downwards collocates (see section 4.2.2.3 for details). We also discard all candidates that occur only in a single MWU template, so that a successful candidate will occur in at least two templates with a frequency of less than the target word.

6.4.3 Problems

There are some initial problems with this procedure. The first one is that the templates we are using for the substitution are automatically identified, so there will be a certain amount of 'noise' from processing mistakes. Since the candidates for multi-word

units are filtered and ranked using a number of general heuristics, any issues that are problematic there will obviously be passed on to the substitution procedure.

The second (and most difficult) problem is that we have no way of knowing whether the multi-word units found are in any way representative of the set of contexts in which the target word occurs. We could end up finding a large number of variations on a more or less fixed expression, whereas other nuances of the word's usage, being less fixed, would not show up in the list. The 'spread' filter, where substitutes from a single MWU template are discarded, does not guard against that. It could be especially problematic for the chains-substitution part.

The solution to the second problem is to link up the substitutions with the words that constitute the MWU with which they are associated. Duplicate entries from similar MWUs (for example *the [dry] season* and *of the [dry] season*) would then be counted only once. We then end up with lists of substitutions attached to the context words from the MWUs to which they apply.

This data was investigated to explore whether a cluster analysis (with the lists of substitutions used as feature vectors) would be able to provide a usable partitioning, but neither of the clustering algorithms used (PAM and AGNES, see Kaufman and Rousseeuw (1990)) yielded any significant structure, as the resulting distance matrix was too sparse. As a next step the mapping was reduced by using only the context word with the lowest frequency, for example *season* from *the [dry] season*. This does in fact reduce the number of resulting mappings down to a manageable size, which can be used without further post-processing.

Occasionally a high-frequency word is the lowest frequency one in a MWU unit, but

such cases can easily be identified as the number of substitutions is far higher than with lexical words. On the other hand, this feature is another manifestation of the continuum between syntactic and semantic classes: a template such as *the [NOUN] of* will find substitutions that (mostly) share the characteristics of a noun with the target word. These words still have more in common with each other than, for example, with words of a different word class. A template containing a specific low-frequency item will usually yield substitutions that share aspects of meaning. Thus even such high-volume substitutions are not a problem from a theoretical point of view, as they fit well into the empirical view of word classes.

The restriction on surface forms has a further limitation: see the example from Winograd (1983) about the *eager/easy* distinction. Looking at data from the written part of the BNC we get a MWU *who are eager to*, and several similar variants. Most of these can also take *easy*, which promptly shows up as a related word. The usage pattern approach only finds *potential* as a related word through the adjective-noun relation. For *easy* a larger number of similar words is identified, also for the same relation. Those words do not include *eager*.

6.4.4 Case Studies

In order to be able to compare different procedures, we have chosen the same case studies and set-up as for the usage pattern approach. In the following sections are the substitutions found for the words by the MWU-Subs procedure described above. The results are shown by context word, with the frequency of occurrence of the substitute associated with the context word.

The full results are also included in appendix C.

6.4.4.1 'Africa'

As we do not reduce phrases to head words, the results for *Africa* are substantially better than they were for the usage pattern substitution.

BBC Africa (409), African (17)

East Africa (154), Asian (3), Berlin (3), Europe (3), Europeans (3)

North Africa(1325), African (23), America (19)

Our Africa (201), African (3)

South Africa (516), Korea (230), America (30), Africans (16), Asia (15), Wales (14), Korean (12), Georgia (7), African (5), Pacific (5), co-operation (5), Pole (4), and (4), Carolina (3), India (3), Koreans (3), Williamson (3), tend (3)

Southern Africa (1532), African (18), AFrica (3)

West Africa (444), African (5), Europe (3)

Some other names of continents are identified, and also names of states that share the same modifiers (*Carolina*, *Korea*). Here we can also see the influence of the BBC corpus, which is responsible for the context word *Our* from the by-line in correspondents' reports.

6.4.4.2 'Germany'

For *Germany* we have a similar picture with the context words *East* and *West*: here we get entities like continents (*Asia*) and cities (*Berlin*, *Beirut*, *Jerusalem*). Also islands (*Indies*, *Timor*) and other areas (*(West) Bank*, *Yorkshire*).

The context word *called* provides other countries and names of groups: *Muslims*, *miners*, *residents* for example. An example frame is *called on Germany to*, where *Germany* is used to refer to a group of people, namely politicians or decision-makers. The substitutions thus are groups of people who are in a position to make decisions, and this feature would be needed to supplement the more general label [GROUP] in a possible abstract representation of this frame as *called on [GROUP] to*.

With *should* (as in *of Germany should be*, a fairly unrestricted MWU) we get a whole range of substitutes, from *people* and *action* to *drugs* and *who*.

Finally, *united* has a much more limited set, *Germany* (639), *front* (25), *campaign* (4), and *approach* (3). With such diverse results it might be necessary to pool the substitutes and treat those that occur only with one context word as suspicious, whereas those which are found with several different context words appear to be more reliable in the sense that they conform more to expectations.

6.4.4.3 'Monday'

Monday has a reasonably large number of candidates; most context words provide a number of other days of the week as substitute candidates. *On* (with an initial capital) gives some names of months, as well as unrelated words such as *paper*. Next, *after* gives

days and events which have a duration (*bail*, *holiday*, *course*, *penalties*, *time*). *Afternoon* has some days of the week plus the determiners *this*, *the*, and *all*.

Overall many other days of the week are identified as similar across various context words, which suggests that it would be useful to summarise the output with the number of different contexts they appear in. Some context words are clearly to unspecific, and occur in multi-word units with a large number of words that are not really in any semantic relation with the target word. For *Monday* we have the context words *first* and *on* which are of limited use. *Morning* on the other hand lists a fair amount of words such as *coffee* and *whole*, but the top seven candidates are all days of the week.

If we summarise the output of the MWU substitution, we can discard all spurious results that are due to unspecific words. We simply take all lists retrieved (19 in this case), and use them to create a frequency list. Words which occur with several context words end up with a higher frequency count. Single occurrences are discarded. For *Monday* we get the following list:

19	Monday	2	board
13	Thursday	2	display
12	Saturday	2	him
12	Tuesday	2	this
11	Friday	2	that
11	Sunday	2	tomorrow
11	Wednesday	2	trade
4	all	2	them
3	the	2	a
2	view		

As we can see, the most frequent substitutes are all days of the week; if we set a threshold to define a cut-off point (e.g. less than half the frequency of the previous

entry in the list) we can effectively get rid of all words which are not member of this closed set.

6.4.4.4 'Smith'

Smith has two context words only with this set-up: *Peter*, and *said*. All of them come up with other surnames, which is a very satisfactory result.

6.4.4.5 'brown'

Brown only provides a small set of data, but unlike the usage patterns there is no case-folding with the MWU substitution, hence the absence of names:

envelope brown (12), sealed (3)

eyes brown (3), beautiful (3), blue (3)

plant brown (3), Scottish (3), coca (3), easiest (3), industrial (3), yam (3), power (3)

These illustrate three very different usages of *brown*, and the number of alternative colour adjectives that could replace it are very limited: eyes can only really be *green*, *red* or *black* (in addition to the ones found), and plants would typically be either *brown* or *green*. However, *green* does not turn up, presumably because plants are green by default and do not need the additional adjective in most situations. Envelopes are white by default, so the colour would not normally be mentioned unless it was non-white.

6.4.4.6 'computer'

Computer also has a large number of substitutes, mainly due to non-specific context words such as *has*, *is*, *said*, *to* and *will*. The more 'useful' context words include:

industry computer, arms, music, petro-chemical, coal, electronics ...

programme computer, BBC, radical, reform, television, big, detailed ...

revolution computer, Thatcher, Chinese, Sandinista, anti-Communist ...

screen computer, large, television

system computer, federal, financial, economic, communist, distribution, regimental ...

Interestingly, computer software is usually referred to as a *program*, which could be a problem with spelling by non-experts. *System* is a very abstract word which does not carry a lot of information by itself.

Summarised we only get *computer* with 24, *television* with 4, and then a number of words with 3 or fewer different context words.

6.4.4.7 'dry'

Here is a selection of the results:

allowed come, dry

area dry, large, complex, different, fertile, small

areas dry, northern, tropical, Kurdish, four, sensitive, troubled, Arab, contaminated, marginal, neighbouring, Muslim, both, built-up, coastal, designated, important, jungle, specific, strategic, vital, arid, country, crowded, deep, industrial, mixed, nearby, patient, separate, slum, small, to, uninhabited, wet

as dry, far, white, late, much, tall

land acquire, building, clear, dry, inherit, leave, own, redistribute, use

out dry, bail, bow, check, climb, dig, drive, miss, throw, weed, buy, drag, drop, hand, jump, knock, lash, leave, look, thrash, back, come, draw, flesh, flush, get, let, ride, search, sell, send, stretch, tease, working ...

season dry, holiday, football, English, off, tourist, close, winter, festive, growing, rainy, Christmas, coming, flood, regular, summer, wet, current, high, hunting, league, lean, monsoon, sailing

summer dry, fine, new

up bring, draw, dry, grow, stand, stay, travel

weight dry, full, political

We can see a range of usages, all reflected in the kinds of candidates identified. The form *dry* can be an adjective describing mainly areas or weather conditions, and we find other adjectives that can describe those. It is also a phrasal verb, used with either *out* or *up* (and also *off*) (Cobuild, 1995), and we find other phrasal verbs used with the same particles.

Applying the summariser we just get *dry* with a frequency of 11, and a handful of other words that occur twice: *clear*, *come*, *draw*, *leave*, *small*, *stand*, *travel*, *wet*, and

working.

6.4.4.8 'plant'

Again, some non-specific context words have been removed (e.g. *was*):

bomb plant, be, defuse

chemical plant, factory, reactions, assistant, complex

life family, plant, your, modern, what, marine, national, normal, British, daily, healthy, our, public, rural, social, this, new, ordinary, private, economic, parliamentary, all, prolonging, supporting, thy

nuclear plant, plants, station, stations, disaster, accidents, equipment, deterrent, facilities, power, presence, warheads, accident, arsenal, artillery, device, experts, fusion, lobby, materials, movement, non-proliferation, proliferation, strategy, test, tests

power plant, struggle, structure, summit, plants, but, rights, vacuum, and, project, station, is, talks, was, failure, relationships, were

We can see the different usages from this list: *plant* as a verb applied to bombs, the industrial meaning of factory or power station, and some biologically related ones (*life*). Here we would have a problem with a summary display, as the units identified are rather diverse and there would be little overlap. It would probably make sense to discard those context words which have a very long list of candidates, as they are always the unspecific ones where the candidates are far too general to make sense. They have already been removed from the list above; the full set is given in appendix C.

6.4.5 Evaluation

Because of the more restricted contexts, the results tend to be better than with the usage pattern approach. Some higher frequency context words indicate that there are multi-word units which are more like empty templates which act as 'syntactic glue' between content words, and are thus not relevant for carrying meaning. But they can sometimes be filtered out when all the context words of a candidate are considered.

The multi-word unit substitution required even less a priori linguistic knowledge than the usage pattern substitution, which relied heavily on traditional syntactic categories. Despite this lack of information that the algorithm has available the results are generally better, though it has been observed for some words that only few substitutes are found, or even none at all. Candidates can be identified only if multi-word units have been found for a word form. Again, a comprehensive analysis of a corpus is required to gather quantitative information, i.e. what degree of coverage of a whole text can be achieved with this method.

6.5 Meaning: Summary and Evaluation

Many computational models of word meaning (e.g. Latent Semantic Indexing, see Deerwester *et al.* 1990) postulate a (Euclidean) semantic space in which words are arranged, with closely related words being spatially close as well. However, when we view similarity of meaning as similarity of shared contexts, we quickly encounter a serious problem, which incidentally is also a major headache for multi-dimensional approaches. Many words (or rather: most words) do not share any contexts at all, and are thus completely unrelated. It does not make sense to ask which of *soup* or *those* is more similar in

meaning to *with*, but a semantic space approach would give a definitive answer to that meaningless question. As a consequence, such approaches suffer from the high dimensionality: each word introduces a further dimension, and thus sparse data, since most of that enormous multi-dimensional space is in fact empty.

As could be expected, a collocational overlap analysis between those three (random) words shows no interpretable structure at all. Other methods would not apply, as the words are unlikely to share any syntactic or phraseological contexts, though it could be imagined that *soup* and *those* might do.

While we are not generally worried about cognitive adequacy (which at present amounts largely to speculation on the workings of the mind), it seems obvious that such an inefficient model is inadequate to the description of semantic relations between words. As an alternative we postulate a more localised model which consists of nodes (words) and vertices between those nodes. Vertices have a particular length which corresponds to the semantic distance between the two connected nodes. Most nodes will have only a small number of vertices, and the absence of a link between two nodes simply means that there is no semantic relationship between them.

This semantic network can be populated incrementally, by adding new words to it which link to the existing ones. The shared-context methods provides us with a consistent way of describing the similarities between a target word and its neighbours in this semantic space. An example of such a network is shown in figure 6.5.

One initial doubt about semantics was that it was completely outside the domain of empirical analysis (Sampson, 2001). However, the three methods described in this chapter clearly show that we can go some way towards uncovering the relationships

between words in an objective way. There is still a lot of refining work to do in order to improve the results, most importantly the inclusion of units of meaning larger than the single word. But even with the current restrictions good results can be achieved.

It is worth stressing at this point that the methods described here somewhat mirror current research in computational linguistics (in the case of the usage pattern substitution), but that they have been implemented in a much smaller context. The current project is predominantly meant as a set of exploratory case studies, which leave space for improvement. At the same time, we have emphasised empirical principles and avoided preconceptions based on existing theories.

In conclusion we can safely say that the statement of Stubbs (2001) on the structure of the vocabulary can be taken as an axiom on which we can successfully base statistical procedures to help us explore the relationships between individual words. With minimal effort we have achieved reasonable results, so it should be possible to develop the methods further and improve our description of the semantic regularities inherent in lexis.

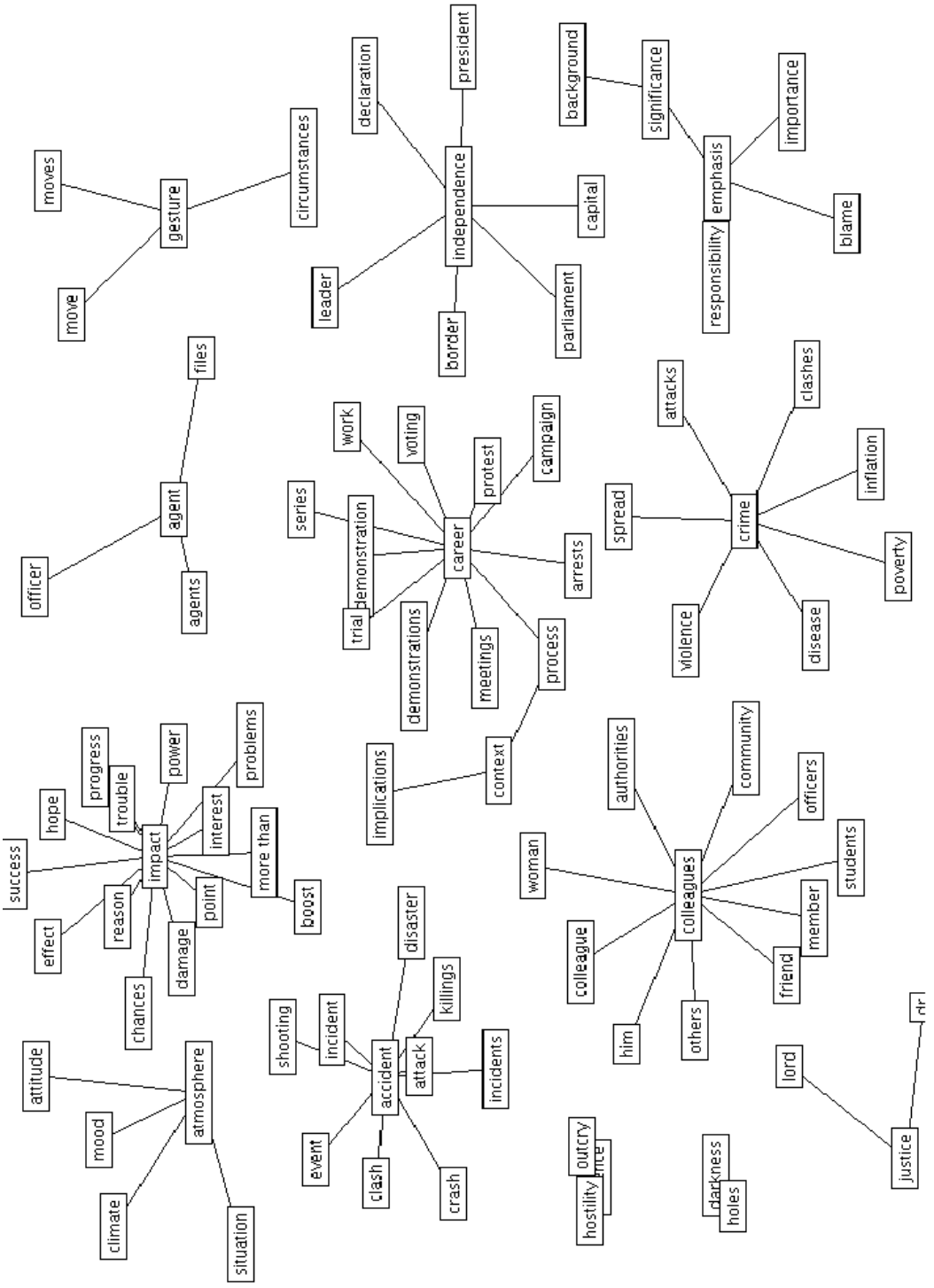

Figure 6.5: An example of a semantic network derived through the usage pattern substitution method

CHAPTER 7

CONCLUSION

In this final chapter we will first summarise the overall results of the research presented in this project. We will then re-visit the project aims and discuss whether they have been achieved, before drawing a number of conclusions. We will end the chapter with a list of topics for future work, partly drawn from the project aims, partly from the conclusions.

7.1 Summary

7.1.1 Research Context

We started off with a brief description of the research background: the development of empirical methods in linguistics in the American and British traditions. The description focused on *discovery procedures*, which aimed to identify linguistic units without any human intervention. Mainly owing to a lack of computational resources or to insufficient data these procedures failed to convince the mainstream research community

that empirical approaches were the way forward, and so they remained largely dormant until about 20 years ago, when statistical methods were introduced through language engineering tasks such as speech recognition.

7.1.2 Methodology

The methods to be used in this project were to be empirical, building on first principles as far as possible. All procedures were to be applied to corpus data, whose properties were discussed. Since it is not (yet?) possible to derive units below word level in a satisfactory way, we presupposed the existence of words as a fundamental unit of language.

Some further features of language also have to be assumed in order to be able to take a 'short-cut', looking at higher levels of language before the lower levels have been comprehensively analysed. Those features include word classes and basic phrases. While the analytic procedures make use of word classes, they do not depend on their actual form, and should also work if and when the word class and phrase structure system has been revised in line with empirical principles.

So some compromises had to be made in order for us to look at 'interesting' parts of language, rather than having to restrict ourselves to trying to discover morphemes from phonemes or graphemes. However, such compromises do not invalidate the overall research, as long as the results are not tied too closely to the tags and phrases used.

7.1.3 Lexis

Looking at lexis, we were able to gather a whole range of useful parameters which describe properties of lexical items or words. These parameters include frequency information, the degree of influence a word exerts on its environment, and the distribution of different inflected forms of a lemma and of tense/aspect/voice (of verbs).

Collocation was identified as a widely used procedure, but one which is usually underspecified: in other words, too often publications do not make clear what choices have been made during the procedure, which makes it impossible to replicate studies. Most parameter settings have a profound influence on the result. Leaving aside some fundamental issues arising from the skewed frequency distribution of words in general (few words are extremely frequent, and most words are very rare), we concluded that the settings depend on the purpose of the analysis. Collocation can thus be defined as an *meta-procedure*, or a template which needs to be instantiated with a set of parameters to yield the procedure for calculating the collocates of a node word.

On the borderline between lexis and syntax we looked at different ways of identifying multi-word units, lying somewhere between single words and phrases. We again have the problem of multiple procedures all producing slightly different results, and in the absence of an *a priori* definition (which would violate the empirical validity of the approach) we cannot easily evaluate the quality of the outcomes, let alone decide which procedure is the correct one. We also touched on using multi-word units as a way into the grammatical description of a sentence: rejecting the presupposed hierarchical nature of a sentence, we can instead describe an utterance through an overlapping sequence of multi-word units derived from a corpus. A sentence would then be made up

of prefabricated chunks which can overlap or be appended end-to-end at a point that happens to coincide with phrase boundaries in traditional constituency grammar.

7.1.4 Grammar

Continuing with grammar, we developed collocation further by modifying one of the algorithms for identifying multi-word units. Colligation was then defined as a sequence of units, either lexical or grammatical, which frequently surrounds the node word. For this we used a traditional phrase structure approach in order to retrieve syntactic units. Other features that might influence the usage of a word (e.g. semantic classes/preferences) could not be considered at this stage. A lot of further work is required to make proper use of colligation, which in the past was rather neglected compared to the (easier) concept of collocation.

Usage patterns were introduced as a further way to describe how words interact at a syntactic level. We again had to forego 'proper' empirical methodology, since we had to use traditional methods of describing the grammatical relationships, namely phrase structure grammar and the basic notions of subject/verb/object. Despite that we were able to extract some useful information about the usage of words.

A different, more empirically-based, approach to grammar was then pursued with grammar patterns. Here we still need to accept predefined units (noun phrases, verb groups), but the relationships between these units are then determined through a set of templates or patterns, extracted from corpus data (unlike grammatical rules, which are mostly based on an individual's ideas about grammar). We were able to identify grammar patterns automatically, given an inventory of existing patterns; the results

did largely match a previous description in a learner's dictionary, though we found that there were further patterns which had been left out, presumably because of space restrictions. We also addressed ways of integrating grammar patterns with a related approach, local grammars.

7.1.5 Meaning

Meaning is one aspect of language that is largely beyond the scope of empirical analysis; only lexical semantics can be approached with current methods. Since the vocabulary of a language has a structure which is reflected in (authentic) utterances, we can analyse the latter to gain insights into the former. Here we must look at the shared element of word meanings, since the associations that individuals have with certain words cannot be identified through the large-scale analysis of corpora. The individual aspects of meaning are in the domain of psycholinguistics, which has a different (empirical) methodology to deal with them.

We looked at three methods, one of which investigates the relationships among a (smallish) group of somehow related words, and two which tried to identify similar words to a target word. The first one was based on collocations, and used the co-occurrence statistics of a set of words (chosen from the merged sets of collocates of the words under investigation) to find out which of those words were more closely related than others. The other two used substitution as a way to evaluate how similar two words were.

The substitution of words with an otherwise identical environment was tested using two different definitions of 'environment': first, the usage patterns, where two words

were deemed to be similar if they shared a number of patterns; and second, lexical environments, linked to the multi-word units derived in the previous chapter. Both methods were found to give useful results, though there were also some unexpected outcomes.

7.2 Discussion

In the first chapter (on page 9) we set out three aims which we wanted to achieve by the end of this project:

- To create a product, a hypertext dictionary/grammar that describes the language of a corpus.
- To develop a software system that can produce such a resource for any corpus.
- To establish which procedures can be implemented, and which procedures require further input data (e.g. of human knowledge) that is not available to the computer.

The first of these aims has been only partially achieved. In principle it is now possible to create such a resource, since the second aim, the creation of a software system for extracting it from corpus data has been achieved. However, the computational resources required (especially for the semantic analysis) were prohibitive, and it was not possible to analyse a full corpus within the time-frame available. Instead, individual sample words were processed as proof-of-concept. As the focus was on the development, we used a rather small and outdated computer for processing; so we assume that, given contemporary hardware, processing speeds will be considerably larger, so that the creation of a dictionary/grammar will be only a matter of setting up the system to run on a larger computer.

The data files created by the system are in the popular XML format, and can therefore be used by any number of subsequent applications that would benefit from having the information available. One such application could be a browser that cross-references entries, so that links are created from the list of collocates of a word to the respective entries of the collocates. The collaborative web environments called 'wikis' would be an easy way to provide a front-end for a human user of the database, or a basic interface could be created within the REST paradigm (Fielding, 2000).

The second aim has been fully achieved. We now have an implementation of a variety of procedures discussed in the previous chapters in the programming language Java, which is portable across a wide variety of platforms. Since the computations involved in processing a given word are independent from processing other words, the whole process is suitable for parallel processing, which would further speed up the creation of the required resource.

At present the system is geared towards the analysis of (written) English data, and several components rely on that: the tokeniser, the parts-of-speech tagger, the lemmatiser, and the parser. The usage patterns and grammar patterns also seem more or less specific to the English language, at least in their present form. However, there is no reason why versions of those tools for different languages should not work; even the usage patterns are based on a set of grammatical relations which exist in other languages as well.

Some procedures will work, but might not produce results in the same way as their English counterparts. The multi-word unit identification relies on the fixed sequences in which words occur; there might be more variability in other languages, and more complex morphology might interfere with the frequency counts. The issue of lemmatisation

is even more dominant in languages other than English. It would thus be interesting to try out the system with corpus data in other languages, and to compare the results with the ones produced for English data.

The third aim, determining whether a given procedure can be automated or not, has also been achieved. While this was easy for those procedures which simply gather distributional information, even those which were assumed to be impossible to automate (e.g. grammar patterns) have been implemented with success. Admittedly, the procedures chosen were mainly ones which were assumed to be suitable. Nevertheless we have shown that we can achieve a broad description of the lexical properties, grammatical features, and semantic aspects of a word by using fully automatic methods which do not require any human intervention at any stage.

The main problem with procedures that extract data from a corpus is that the algorithm does not know which of the outcomes are relevant or useful, and which are merely chance. With some basic assumptions about the nature of language and some (arbitrary) filters based on frequency differentials, we were able to reduce the amount of data to a manageable size, while remaining confident that the degree of coverage was high enough.

Overall, we can say that the aims have been achieved wherever possible, and that those objectives still to be fulfilled will not pose any insurmountable problems; it is only the demand for computational power that has stopped us from performing a full-scale analysis.

7.3 Conclusions

The thesis underlying this project was phrased as follows:

> The description of (a sample of) language can be automated to a high degree. Through large-scale comprehensive analysis of linguistic phenomena new insights can be gained which would not be possible with small-scale manual work. Thus automated analysis not only provides a quantitative gain, but also a qualitative one.

It has been shown in the previous chapters that the analysis of language can indeed be automated, that we can gain new insights into the nature of language by investigating the results of the automated procedures, and that the advance in knowledge is not purely a quantitative one.

From the work carried out throughout this project we can draw four major conclusions:

1. Empirical methods constitute a valid approach to the study of language
2. Linguistic units are inherently vague and require re-definition
3. Large-scale comparisons of linguistic phenomena are useful for the description of language
4. The precise implementation of analytical procedures cannot be separated from their application

In the following four sections we will discuss these conclusions in more detail.

7.3.1 The Empirical Process

Post-war linguistics was dominated by the shift away from empirical methods to intuition as a guiding principle. Even today the latest theories (Archangeli and Langendoen, 1997) invent their own data on which to build up the descriptive formalism. The main reason for this shift was the lack of results: while there was a lot of optimism about discovery procedures, they failed to deliver, mainly owing to factors that were not relevant from a methodological point of view (lack of data and computing power). However, as more powerful computers have become available, and with them electronic corpora, these factors are no longer inhibiting empirical research. Since the mid-1980s, an increasing body of research in corpus linguistics has been built up. Admittedly, not all of this research was truly empirical: much of it was just looking for corroborative evidence for categories derived in non-empirical ways.

The parallel development of two strands of linguistics, empirical and non-empirical, has created a difficult situation for the empiricists: it is now assumed that language description has to employ certain categories and make use of certain larger units and structures, all of which have been created through introspective analysis. As will be described in the next section, these units do not necessarily have any empirical basis, and thus it is not possible to replicate the linguistic structures with fully empirical methods. This mismatch of introspectively created structures and empirically generated ones, together with the dominance of the non-empirical view of linguistics, makes it extremely hard for empiricists to argue for the validity of their results.

However, if we reject non-empirical approaches and try not to be influenced by preconceptions drawn from traditional linguistics, it is clear that we can produce results that are coherent and internally consistent and provide a valid description of language—albeit one that appears unusual, as it diverges from what linguists are used to because of their traditional training. It is no surprise that some support for the empirical approach has come from other disciplines, such as computing and engineering, where it is more commonly accepted that the real world does not always fit neatly into the elaborate and aesthetically pleasing theories created by armchair-practitioners.

The research described in the context of this project has created plausible results, which can be evaluated in the light of what we know about the general properties of language, such as statistical regularities. Just as Danielsson (2001) applies Zipf's law to evaluate the outcome of her procedures to find units of meaning, we can appeal to similar principles and regularities to justify our results. In fact, using principles that have been shown to have universal qualities (such as Zipf's law, which applies to a large variety of phenomena in unrelated subject areas) provides much stronger evidence than the resemblance to structures that have been made up by individuals with particular ideas about how language should be organised.

7.3.2 Units of Language

Continuing the argument from the previous section, the main issue of the empirical/intuitive divide is the incompatibility of results. The incompatibility arises from the definition of the fundamental units of analysis, on which any subsequent work is based. Traditional grammar has always concentrated on written language, and on those sentences that follow prescribed grammatical rules. It is therefore heavily influenced by

conventions of spelling and prescriptive notions of what constitutes 'proper' language. The basic unit here is the word, since words are separated by white space in writing. Morphemes have been postulated to deal with the observation that many words have parts in common, and that there is some correlation between those (shared) parts and elements of meaning of the word.

In (lexical) semantics, words are treated as units of meaning (though morphemes are commonly defined as the smallest unit of language that carries meaning); but as we have seen, words on their own have a meaning potential that is only realised in a context. A word in isolation has no meaning. When adding the minimally required context to the word, we end up with larger units, above the level of words but usually below the level of the phrase, the next unit postulated in traditional grammar. The units of meaning that can be identified empirically do not map on to any existing unit, which makes it hard to evaluate and compare the outcomes. On a similar issue Esser (2000) has suggested an alternative empirical definition of the *linguistic sign*, which improves on the original definition by de Saussure (1916).

The problem is that the existing units are so deeply ingrained in the linguistic tradition that it is very difficult to change preconceptions. Language is undoubtedly a symbolic system, but it is based on a sub-symbolic one: the human brain. At some stage during language processing the sub-symbolic activation patterns within parts of the brain turn into symbolic entities. Furthermore, language is not static, but adapts to an ever-changing environment. For these two reasons it is unlikely that the resulting symbolic system will be consistent and logical, as assumed by the traditional paradigm.

What is needed now is a proper empirical foundation on which to base the description of language. Linguistic units need to be redefined according to general principles

shown to be valid across disciplines. But most importantly, linguists need to be prepared to let go of their preconceptions with regard to linguistic units.

7.3.3 Large-scale Comparisons

Owing to computational restrictions, we have been able to perform only a small number of large-scale comparisons on features that could be collected without too much effort. However, even those few case studies have shown that we can find out a lot about the way language is organised. Since language is ultimately based on a sub-symbolic system, it will have to follow certain organisational principles from physics or biology. And only by investigating large sets of data across many different sources can we identify how such principles are applied.

A basic difficulty in this task lies in the definition of units described in the previous section: we have been looking at words, when the appropriate unit would probably have been larger. As a consequence the results might not show the 'true' distribution, though the whole analysis then turns into a knot of Gordian proportions, where it becomes increasingly difficult to find the starting point. The only way forward will be a step-by-step refinement of procedures, using flexible definitions of units (such as multi-word units with added/omitted/transposed elements), until a set of consistent results has been obtained.

It is not possible to learn about the organisation and structure of language by looking only at small samples; instead we need to take a broad view to discover patterns in the data.

7.3.4 Theory and Application

Linguistic research does not exist in a void; instead it needs to fulfill a variety of needs. They include the description of language for teaching purposes, for applications such as speech recognition or machine translation, and also for general research into the nature of language. It is not possible to satisfy all these demands with a single approach.

One prominent example is collocation. It does not make any sense to research collocation in the abstract, since different settings for the parameters involved produce wildly different results. Comparing different significance functions is meaningless without a) considering other parameters and b) having some notion of the desired outcome. Otherwise one might as well use a random number generator to assign a significance value to collocates.

However, most researchers still use collocations without being aware of the randomness of their results. In general there seems to be too much trust in the software which performs the analysis, and too little information for the user about the way it is done. Software authors should make all settings explicit, and users should enquire about them. Without parameter settings being clarified, research is not replicable and thus of limited validity.

When researching the effects that different parameter settings have on the outcomes of a procedure such as collocation, abstract notions will need to be considered, so that particular choices can be set out in terms of describing the form of the outcomes. For example, 'these settings favour words of medium frequency which tend to occur mostly on the left-hand side of the node word'. That would describe a setting suitable for the

analysis of adjectives that modify a particular noun.

7.4 Future Work

In this section we will briefly describe work that we planned to do but could not in the time-frame available, and we will also discuss some further questions raised by this current project. Both will provide pathways for continuing the research begun here, which after all was meant to be only a feasibility study. Having shown that fully automatic empirical analysis of language data is indeed feasible, we will now outline a programme for future work.

7.4.1 Planned Work

One of the aims of this project was to create a linguistic resource that could be consulted by both humans and machines. While the latter objective has been achieved in principle through the creation of a set of XML files containing the extracted information in a form that can be easily accessed by other applications, it is less satisfactory for a human user. One desirable application would thus be a *lexicon browser* that makes the data more accessible. This is a trivial task whose main effort would be to design an appropriate graphical user interface for easy navigation. The easiest approach would probably be to create a module that resides on a web-server and creates dynamic web-pages from the database. Using stylesheets it is possible to control the amount of information that is actually displayed, so that details can be hidden by default. By selecting a link, those details can then be revealed by manipulating the web-page within the browser. Alternatively, a simple stand-alone application could be created that would operate directly on

the data files without the additional layer of a web-server.

Before such an application could be used, however, the database would need to be complete. This would require a (lengthy) processing run on a reasonably powerful computing system. The software could perhaps optimised to reduce the run-time requirements, since it was originally designed to be flexible. Flexibility, while necessary during the research and development stage, is usually bad for performance. Some optimisations were already needed to solve the more demanding tasks, and with careful profiling it ought to be possible to improve further the run-time performance of the system.

Several procedures have been implemented as prototypes, but were left out of the overall system due to lack of time and space. These procedures could for example perform further analysis of semantic behaviour, or could attempt to find generalisations based on (pre-determined) semantic classes. It was originally planned to provide the system with a way of identifying such semantic classes as 'day of week', 'colour', 'piece of furniture', etc. However, such a procedure would have required the creation of an extensive list of words with appropriate classes, which would not have been in the empirical spirit of the project. It might nevertheless be interesting to add such a component to see whether collocational analysis or perhaps even usage patterns could benefit from it.

Finally, integration with existing frameworks would have been desirable in order to facilitate the re-use of existing linguistic resources, or the easier use of resources newly created by other researchers. As an example, it would have been good to create automata for multi-word units in a form compatible with the INTEX system. While this is not a problem in principle, it was postponed because of lack of time; furthermore, we

initially planned to use INTEX for parts of this project, but in the end did not do so.

7.4.2 Further Issues

Most research throws up more questions than it answers, and this project is no exception. Especially when challenging traditional notions like units of analysis potentially invalidates much previous research. Apart from investigating further procedures for extracting linguistic information, we need to revisit previous methods with different settings. Ideally a corpus should be annotated with multi-word units (similar to what Danielsson (2001) did with her units of meaning) and then the analysis should be re-run, providing different sets of frequencies, gravities, collocates, and even second-level multi-word units. The semantic algorithms might in fact perform better, since they would no longer be limited to single words.

But this would not work for those methods which presuppose the traditional system of word classes and phrase structures. Different ways would have to be found to describe usage patterns and grammar patterns. The word class system would also require re-definition when applied to multi-word units; eventually we might have to ask whether it was applicable in the first place, when multi-word units no longer fitted into such clear-cut 'slots' as single words. We might need a completely new and different model of grammar.

On the other hand, several more established methods of description could be explored, for example semantic frames (as described in Fillmore's FrameNet project) or the process types of systemic functional grammar. In a similar way to pattern grammar, it should be possible to identify the distributional properties of frames and process

types.

In conclusion, recent advances in natural language processing have made it practicable to explore linguistic issues on a much larger scale than has been possible through manual work. By looking at phenomena on this larger scale we not only advance quantitatively, but we also gain new insights on a qualitative level. The near future will be an exciting time for computational corpus linguistics.

CHAPTER A

PART OF SPEECH LABELS

The part of speech labels used are a slightly modified version of the family of tags used for the Brown corpus and the Penn treebank. This is the most widely used set of tags for English data.

-	punctuation
,	punctuation
;	punctuation
:	punctuation
!	punctuation
?	punctuation
???	unknown word
.	punctuation
...	punctuation
"	punctuation
(	punctuation
)	punctuation
BE	*be*
BED	*were*
BEDZ	*was*
BEG	*being*
BEM	*am*
BEN	*been*
BER	*are*
BEZ	*is*
CC	coordinating conjunction
CD	cardinal number
CS	conjunction

DO	*do*
DOD	*did*
DOG	*doing*
DON	*done*
DOZ	*does*
DT	determiner
EX	existential *there*
FW	foreign word
HV	*have*
HVD	*had* (past tense)
HVG	*having*
HVN	*had* (pas participle)
HVZ	*has*
IN	preposition, subordinating conjunction
JJ	adjective
JJR	comparative adjective
JJS	superlative adjective
MD	modal
NN	noun
NNS	plural noun
NP	proper noun
NPS	plural proper noun
OD	ordinal number
PDT	pre-determiner
PN	pronoun
POS	possessive *'s*
PP	personal pronoun
PP$	possessive pronoun
PPX	reflexive pronoun
RB	adverb
RBR	comparative adverb
RBS	superlative adverb
RP	particle
SYM	symbol
TO	infinitive marker *to*
UH	interjection
VB	base form of verb
VBD	past tense verb
VBG	continuous form
VBN	past participle
VBZ	3rd person singular verb
WDT	wh-determiner
WP	wh-pronoun
WP$	possessive wh-pronoun
WRB	wh-adverb
XNOT	*not* and contracted forms

CHAPTER B

XML SAMPLE OUTPUT

This is an example of a complete output file as generated by the system. The word form is *dry*, and the corpus used is the BBC corpus.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<dict_entry>
    <wordform>dry</wordform>
    <corpus>BBC</corpus>
    <state value="10"/>
    <basic_statistics>
        <frequency>317</frequency>
        <freqPer10K>0.17064850090944347</freqPer10K>
        <freqBandTotal>12</freqBandTotal>
        <inflections>
            <entry>
                <wordform>dried</wordform>
                <freqband>13</freqband>
            </entry>
            <entry>
                <wordform>drier</wordform>
                <freqband>16</freqband>
            </entry>
            <entry>
                <wordform>dries</wordform>
                <freqband>16</freqband>
            </entry>
            <entry>
                <wordform>driest</wordform>
                <freqband>20</freqband>
            </entry>
            <entry>
                <wordform>dry</wordform>
                <freqband>12</freqband>
            </entry>
            <entry>
                <wordform>dryer</wordform>
                <freqband>18</freqband>
            </entry>
            <entry>
                <wordform>drying</wordform>
                <freqband>14</freqband>
            </entry>
        </inflections>
    </basic_statistics>
    <lexical_gravity>
        <entropy offset="-15">7.1407404673674595</entropy>
        <entropy offset="-14">6.923764645563695</entropy>
        <entropy offset="-13">7.0027199903270345</entropy>
        <entropy offset="-12">6.751522259423432</entropy>
        <entropy offset="-11">6.918737096957897</entropy>
        <entropy offset="-10">7.025593839792786</entropy>
        <entropy offset="-9">6.823655920584142</entropy>
        <entropy offset="-8">7.06685186707272 7</entropy>
        <entropy offset="-7">7.012997913390412</entropy>
        <entropy offset="-6">6.902290483641251</entropy>
        <entropy offset="-5">6.780225603446954</entropy>
        <entropy offset="-4">6.907675731115034</entropy>
        <entropy offset="-3">6.980099242337572</entropy>
        <entropy offset="-2">6.6048251464403</entropy>
        <entropy offset="-1">5.595748586649671</entropy>
        <entropy offset="0">-0.0</entropy>
        <entropy offset="1">5.821192888163118</entropy>
        <entropy offset="2">5.949795483093489</entropy>
        <entropy offset="3">6.698875231142313</entropy>
        <entropy offset="4">6.9233769068177375</entropy>
        <entropy offset="5">6.9306540032604875</entropy>
        <entropy offset="6">6.869042581942047</entropy>
        <entropy offset="7">7.093376284538996</entropy>
        <entropy offset="8">7.066863251369634</entropy>
        <entropy offset="9">6.968727452050726</entropy>
        <entropy offset="10">6.954095436396 02</entropy>
        <entropy offset="11">6.767548062064373</entropy>
        <entropy offset="12">7.222753163542572</entropy>
        <entropy offset="13">6.666332878801535</entropy>
        <entropy offset="14">6.953416930159206</entropy>
        <entropy offset="15">6.950965874785485</entropy>
    </lexical_gravity>
    <span>
        <left>2</left>
        <right>3</right>
    </span>
    <collocations>
        <parameter>
            <name>window</name>
            <value>rectangular</value>
        </parameter>
        <parameter>
            <name>significance</name>
            <value>rawFreq</value>
        </parameter>
        <parameter>
            <name>threshold</name>
            <value>3</value>
        </parameter>
        <collocate id="0">
            <form>bread</form>
            <freqband>12</freqband>
            <frequency>8</frequency>
        </collocate>
        <collocate id="1">
            <form>begging</form>
            <freqband>14</freqband>
            <frequency>7</frequency>
        </collocate>
```

```
    <collocate id="2">
        <form>climates</form>
        <freqband>15</freqband>
        <frequency>4</frequency>
    </collocate>
    <collocate id="3">
        <form>consecutive</form>
        <freqband>13</freqband>
        <frequency>4</frequency>
    </collocate>
    <collocate id="4">
        <form>dangerously</form>
        <freqband>13</freqband>
        <frequency>4</frequency>
    </collocate>
    <collocate id="5">
        <form>ink</form>
        <freqband>15</freqband>
        <frequency>4</frequency>
    </collocate>
    <collocate id="6">
        <form>tinder</form>
        <freqband>18</freqband>
        <frequency>4</frequency>
    </collocate>
    <collocate id="7">
        <form>wet</form>
        <freqband>13</freqband>
        <frequency>4</frequency>
    </collocate>
    <collocate id="8">
        <form>wood</form>
        <freqband>12</freqband>
        <frequency>4</frequency>
    </collocate>
    <collocate id="9">
        <form>facts</form>
        <freqband>12</freqband>
        <frequency>3</frequency>
    </collocate>
    <collocate id="10">
        <form>farming</form>
        <freqband>12</freqband>
        <frequency>3</frequency>
    </collocate>
    <collocate id="11">
        <form>hectare</form>
        <freqband>14</freqband>
        <frequency>3</frequency>
    </collocate>
    <collocate id="12">
        <form>impersonal</form>
        <freqband>17</freqband>
        <frequency>3</frequency>
    </collocate>
    <collocate id="13">
        <form>spell</form>
        <freqband>12</freqband>
        <frequency>3</frequency>
    </collocate>
    <collocate id="14">
        <form>summers</form>
        <freqband>17</freqband>
        <frequency>3</frequency>
    </collocate>
    <collocate id="15">
        <form>woodland</form>
        <freqband>15</freqband>
        <frequency>3</frequency>
    </collocate>
</collocations>
<frames>
    <frame id="0">
        <frequency>2</frequency>
        <phrase>and a
            <node>dry</node> summer last year
        </phrase>
    </frame>
    <frame id="1">
        <frequency>2</frequency>
        <phrase>hadn't always been considered too
            <node>dry</node> for
        </phrase>
    </frame>
    <frame id="2">
        <frequency>2</frequency>
        <phrase>have been completely
            <node>dry</node> since June
        </phrase>
    </frame>
</frames>
<chains>
    <concSize>275</concSize>
    <chain id="0">
        <frequency>20</frequency>
        <phrase>would eat
            <node>dry</node> bread before begging
        </phrase>
    </chain>
    <chain id="1">
        <frequency>17</frequency>
        <phrase>the
            <node>dry</node> season
        </phrase>
    </chain>
    <chain id="2">
        <frequency>15</frequency>
        <phrase>eat
            <node>dry</node> bread before begging
        </phrase>
    </chain>
    <chain id="3">
        <frequency>15</frequency>
        <phrase>would eat
            <node>dry</node> bread before
        </phrase>
    </chain>
    <chain id="4">
        <frequency>12</frequency>
        <phrase>of the
            <node>dry</node> season
        </phrase>
    </chain>
    <chain id="5">
        <frequency>10</frequency>
        <phrase>Barbados hadn't always been considered
            too
            <node>dry</node>
        </phrase>
    </chain>
    <chain id="6">
        <frequency>10</frequency>
        <phrase>Pakistanis would eat
            <node>dry</node> bread before begging
        </phrase>
    </chain>
    <chain id="7">
        <frequency>10</frequency>
        <phrase>
            <node>dry</node> bread before begging
        </phrase>
    </chain>
    <chain id="8">
        <frequency>10</frequency>
        <phrase>
            <node>dry</node> day the snails retire
            beneath the
        </phrase>
    </chain>
    <chain id="9">
        <frequency>10</frequency>
        <phrase>
            <node>dry</node> it and it is becoming
            more
        </phrase>
    </chain>
    <chain id="10">
        <frequency>10</frequency>
        <phrase>average rainfall and a
            <node>dry</node> summer last
        </phrase>
    </chain>
    <chain id="11">
        <frequency>10</frequency>
        <phrase>below average rainfall and a
            <node>dry</node> summer
        </phrase>
    </chain>
    <chain id="12">
        <frequency>10</frequency>
        <phrase>can even maybe climb ships in
            <node>dry</node>
        </phrase>
    </chain>
    <chain id="13">
        <frequency>10</frequency>
```

```
    <phrase>eat
        <node>dry</node> bread before
    </phrase>
</chain>
<chain id="14">
    <frequency>10</frequency>
    <phrase>even maybe climb ships in
        <node>dry</node> docks
    </phrase>
</chain>
<chain id="15">
    <frequency>10</frequency>
    <phrase>factories had also started to
        <node>dry</node> up
    </phrase>
</chain>
<chain id="16">
    <frequency>10</frequency>
    <phrase>for people farming in
        <node>dry</node> areas than
    </phrase>
</chain>
<chain id="17">
    <frequency>10</frequency>
    <phrase>greater for people farming in
        <node>dry</node> areas
    </phrase>
</chain>
<chain id="18">
    <frequency>10</frequency>
    <phrase>hadn't always been considered too
        <node>dry</node> for
    </phrase>
</chain>
<chain id="19">
    <frequency>10</frequency>
    <phrase>in the
        <node>dry</node> season
    </phrase>
</chain>
<chain id="20">
    <frequency>10</frequency>
    <phrase>is now said to be tinder
        <node>dry</node>
    </phrase>
</chain>
<chain id="21">
    <frequency>10</frequency>
    <phrase>is to
        <node>dry</node> it and it is
    </phrase>
</chain>
<chain id="22">
    <frequency>10</frequency>
    <phrase>of below average rainfall and a
        <node>dry</node>
    </phrase>
</chain>
<chain id="23">
    <frequency>10</frequency>
    <phrase>rainfall and a
        <node>dry</node> summer last year
    </phrase>
</chain>
<chain id="24">
    <frequency>10</frequency>
    <phrase>taps have been completely
        <node>dry</node> since June
    </phrase>
</chain>
<chain id="25">
    <frequency>10</frequency>
    <phrase>the factories had also started to
        <node>dry</node>
    </phrase>
</chain>
<chain id="26">
    <frequency>10</frequency>
    <phrase>to
        <node>dry</node> it and it is becoming
    </phrase>
</chain>
<chain id="27">
    <frequency>10</frequency>
    <phrase>water taps have been completely
        <node>dry</node> since
    </phrase>
</chain>
<chain id="28">
    <frequency>10</frequency>
    <phrase>would eat
        <node>dry</node> bread
    </phrase>
</chain>
<chain id="29">
    <frequency>9</frequency>
    <phrase>after fifteen consecutive
        <node>dry</node> days
    </phrase>
</chain>
<chain id="30">
    <frequency>9</frequency>
    <phrase>of the
        <node>dry</node>
    </phrase>
</chain>
<chain id="31">
    <frequency>8</frequency>
    <phrase>Pakistanis would eat
        <node>dry</node> bread before
    </phrase>
</chain>
<chain id="32">
    <frequency>8</frequency>
    <phrase>
        <node>dry</node> day the snails retire
        beneath
    </phrase>
</chain>
<chain id="33">
    <frequency>8</frequency>
    <phrase>
        <node>dry</node> it and it is becoming
    </phrase>
</chain>
<chain id="34">
    <frequency>8</frequency>
    <phrase>
        <node>dry</node> them in the presence of
    </phrase>
</chain>
<chain id="35">
    <frequency>8</frequency>
    <phrase>always been considered too
        <node>dry</node> for
    </phrase>
</chain>
<chain id="36">
    <frequency>8</frequency>
    <phrase>and a
        <node>dry</node> summer last year
    </phrase>
</chain>
<chain id="37">
    <frequency>8</frequency>
    <phrase>average rainfall and a
        <node>dry</node> summer
    </phrase>
</chain>
<chain id="38">
    <frequency>8</frequency>
    <phrase>below average rainfall and a
        <node>dry</node>
    </phrase>
</chain>
<chain id="39">
    <frequency>8</frequency>
    <phrase>even maybe climb ships in
        <node>dry</node>
    </phrase>
</chain>
<chain id="40">
    <frequency>8</frequency>
    <phrase>factories had also started to
        <node>dry</node>
    </phrase>
</chain>
<chain id="41">
    <frequency>8</frequency>
    <phrase>for people farming in
        <node>dry</node> areas
    </phrase>
</chain>
<chain id="42">
    <frequency>8</frequency>
    <phrase>frozen carbon dioxide or
```

```
        <node>dry</node> ice
    </phrase>
</chain>
<chain id="43">
    <frequency>8</frequency>
    <phrase>greater for people farming in
        <node>dry</node>
    </phrase>
</chain>
<chain id="44">
    <frequency>8</frequency>
    <phrase>had also started to
        <node>dry</node> up
    </phrase>
</chain>
<chain id="45">
    <frequency>8</frequency>
    <phrase>hadn't always been considered too
        <node>dry</node>
    </phrase>
</chain>
<chain id="46">
    <frequency>8</frequency>
    <phrase>have been completely
        <node>dry</node> since June
    </phrase>
</chain>
<chain id="47">
    <frequency>8</frequency>
    <phrase>is to
        <node>dry</node> it and it
    </phrase>
</chain>
<chain id="48">
    <frequency>8</frequency>
    <phrase>maybe climb ships in
        <node>dry</node> docks
    </phrase>
</chain>
<chain id="49">
    <frequency>8</frequency>
    <phrase>now said to be tinder
        <node>dry</node>
    </phrase>
</chain>
<chain id="50">
    <frequency>8</frequency>
    <phrase>people farming in
        <node>dry</node> areas than
    </phrase>
</chain>
<chain id="51">
    <frequency>8</frequency>
    <phrase>rainfall and a
        <node>dry</node> summer last
    </phrase>
</chain>
<chain id="52">
    <frequency>8</frequency>
    <phrase>taps have been completely
        <node>dry</node> since
    </phrase>
</chain>
<chain id="53">
    <frequency>8</frequency>
    <phrase>to
        <node>dry</node> it and it is
    </phrase>
</chain>
<chain id="54">
    <frequency>8</frequency>
    <phrase>water taps have been completely
        <node>dry</node>
    </phrase>
</chain>
<chain id="55">
    <frequency>7</frequency>
    <phrase>in the
        <node>dry</node>
    </phrase>
</chain>
<chain id="56">
    <frequency>7</frequency>
    <phrase>to
        <node>dry</node> up
    </phrase>
</chain>
<chain id="57">
    <frequency>6</frequency>
    <phrase>Pakistanis would eat
        <node>dry</node> bread
    </phrase>
</chain>
<chain id="58">
    <frequency>6</frequency>
    <phrase>
        <node>dry</node> day the snails retire
    </phrase>
</chain>
<chain id="59">
    <frequency>6</frequency>
    <phrase>
        <node>dry</node> it and it is
    </phrase>
</chain>
<chain id="60">
    <frequency>6</frequency>
    <phrase>
        <node>dry</node> them in the presence
    </phrase>
</chain>
<chain id="61">
    <frequency>6</frequency>
    <phrase>a
        <node>dry</node> summer last year
    </phrase>
</chain>
<chain id="62">
    <frequency>6</frequency>
    <phrase>after fifteen consecutive
        <node>dry</node>
    </phrase>
</chain>
<chain id="63">
    <frequency>6</frequency>
    <phrase>also started to
        <node>dry</node> up
    </phrase>
</chain>
<chain id="64">
    <frequency>6</frequency>
    <phrase>always been considered too
        <node>dry</node>
    </phrase>
</chain>
<chain id="65">
    <frequency>6</frequency>
    <phrase>and a
        <node>dry</node> summer last
    </phrase>
</chain>
<chain id="66">
    <frequency>6</frequency>
    <phrase>average rainfall and a
        <node>dry</node>
    </phrase>
</chain>
<chain id="67">
    <frequency>6</frequency>
    <phrase>been completely
        <node>dry</node> since June
    </phrase>
</chain>
<chain id="68">
    <frequency>6</frequency>
    <phrase>been considered too
        <node>dry</node> for
    </phrase>
</chain>
<chain id="69">
    <frequency>6</frequency>
    <phrase>carbon dioxide or
        <node>dry</node> ice
    </phrase>
</chain>
<chain id="70">
    <frequency>6</frequency>
    <phrase>climb ships in
        <node>dry</node> docks
    </phrase>
</chain>
<chain id="71">
    <frequency>6</frequency>
    <phrase>during the
        <node>dry</node> season
    </phrase>
```

```
</chain>
<chain id="72">
    <frequency>6</frequency>
    <phrase>farming in
        <node>dry</node> areas than
    </phrase>
</chain>
<chain id="73">
    <frequency>6</frequency>
    <phrase>fifteen consecutive
        <node>dry</node> days
    </phrase>
</chain>
<chain id="74">
    <frequency>6</frequency>
    <phrase>for people farming in
        <node>dry</node>
    </phrase>
</chain>
<chain id="75">
    <frequency>6</frequency>
    <phrase>frozen carbon dioxide or
        <node>dry</node>
    </phrase>
</chain>
<chain id="76">
    <frequency>6</frequency>
    <phrase>had also started to
        <node>dry</node>
    </phrase>
</chain>
<chain id="77">
    <frequency>6</frequency>
    <phrase>have been completely
        <node>dry</node> since
    </phrase>
</chain>
<chain id="78">
    <frequency>6</frequency>
    <phrase>is to
        <node>dry</node> it and
    </phrase>
</chain>
<chain id="79">
    <frequency>6</frequency>
    <phrase>it a very
        <node>dry</node> area
    </phrase>
</chain>
<chain id="80">
    <frequency>6</frequency>
    <phrase>maybe climb ships in
        <node>dry</node>
    </phrase>
</chain>
<chain id="81">
    <frequency>6</frequency>
    <phrase>of the
        <node>dry</node> season has
    </phrase>
</chain>
<chain id="82">
    <frequency>6</frequency>
    <phrase>people farming in
        <node>dry</node> areas
    </phrase>
</chain>
<chain id="83">
    <frequency>6</frequency>
    <phrase>rainfall and a
        <node>dry</node> summer
    </phrase>
</chain>
<chain id="84">
    <frequency>6</frequency>
    <phrase>said to be tinder
        <node>dry</node>
    </phrase>
</chain>
<chain id="85">
    <frequency>6</frequency>
    <phrase>taps have been completely
        <node>dry</node>
    </phrase>
</chain>
<chain id="86">
    <frequency>6</frequency>
    <phrase>the coming
        <node>dry</node> season
    </phrase>
</chain>
<chain id="87">
    <frequency>6</frequency>
    <phrase>the ink
        <node>dry</node> on
    </phrase>
</chain>
<chain id="88">
    <frequency>6</frequency>
    <phrase>the ink
        <node>dry</node> on the
    </phrase>
</chain>
<chain id="89">
    <frequency>6</frequency>
    <phrase>to
        <node>dry</node> it and it
    </phrase>
</chain>
<chain id="90">
    <frequency>6</frequency>
    <phrase>was the ink
        <node>dry</node> on
    </phrase>
</chain>
<chain id="91">
    <frequency>5</frequency>
    <phrase>
        <node>dry</node> bread before
    </phrase>
</chain>
<chain id="92">
    <frequency>5</frequency>
    <phrase>eat
        <node>dry</node> bread
    </phrase>
</chain>
<chain id="93">
    <frequency>5</frequency>
    <phrase>home and
        <node>dry</node>
    </phrase>
</chain>
<chain id="94">
    <frequency>5</frequency>
    <phrase>in
        <node>dry</node> areas
    </phrase>
</chain>
<chain id="95">
    <frequency>5</frequency>
    <phrase>would eat
        <node>dry</node>
    </phrase>
</chain>
<chain id="96">
    <frequency>4</frequency>
    <phrase>Pakistanis would eat
        <node>dry</node>
    </phrase>
</chain>
<chain id="97">
    <frequency>4</frequency>
    <phrase>
        <node>dry</node> day the snails
    </phrase>
</chain>
<chain id="98">
    <frequency>4</frequency>
    <phrase>
        <node>dry</node> it and it
    </phrase>
</chain>
<chain id="99">
    <frequency>4</frequency>
    <phrase>
        <node>dry</node> on the
    </phrase>
</chain>
<chain id="100">
    <frequency>4</frequency>
    <phrase>
        <node>dry</node> summer last year
    </phrase>
</chain>
<chain id="101">
```

```
    <frequency>4</frequency>
    <phrase>
        <node>dry</node> them in the
    </phrase>
</chain>
<chain id="102">
    <frequency>4</frequency>
    <phrase>a
        <node>dry</node> summer last
    </phrase>
</chain>
<chain id="103">
    <frequency>4</frequency>
    <phrase>a very
        <node>dry</node> area
    </phrase>
</chain>
<chain id="104">
    <frequency>4</frequency>
    <phrase>ability to
        <node>dry</node> out
    </phrase>
</chain>
<chain id="105">
    <frequency>4</frequency>
    <phrase>also started to
        <node>dry</node>
    </phrase>
</chain>
<chain id="106">
    <frequency>4</frequency>
    <phrase>and a
        <node>dry</node> summer
    </phrase>
</chain>
<chain id="107">
    <frequency>4</frequency>
    <phrase>been completely
        <node>dry</node> since
    </phrase>
</chain>
<chain id="108">
    <frequency>4</frequency>
    <phrase>been considered too
        <node>dry</node>
    </phrase>
</chain>
<chain id="109">
    <frequency>4</frequency>
    <phrase>carbon dioxide or
        <node>dry</node>
    </phrase>
</chain>
<chain id="110">
    <frequency>4</frequency>
    <phrase>climb ships in
        <node>dry</node>
    </phrase>
</chain>
<chain id="111">
    <frequency>4</frequency>
    <phrase>coming
        <node>dry</node> season
    </phrase>
</chain>
<chain id="112">
    <frequency>4</frequency>
    <phrase>completely
        <node>dry</node> since June
    </phrase>
</chain>
<chain id="113">
    <frequency>4</frequency>
    <phrase>considered too
        <node>dry</node> for
    </phrase>
</chain>
<chain id="114">
    <frequency>4</frequency>
    <phrase>dioxide or
        <node>dry</node> ice
    </phrase>
</chain>
<chain id="115">
    <frequency>4</frequency>
    <phrase>farming in
        <node>dry</node> areas
    </phrase>
</chain>
<chain id="116">
    <frequency>4</frequency>
    <phrase>fruit inside is
        <node>dry</node>
    </phrase>
</chain>
<chain id="117">
    <frequency>4</frequency>
    <phrase>have been completely
        <node>dry</node>
    </phrase>
</chain>
<chain id="118">
    <frequency>4</frequency>
    <phrase>home and
        <node>dry</node> in
    </phrase>
</chain>
<chain id="119">
    <frequency>4</frequency>
    <phrase>in
        <node>dry</node> areas than
    </phrase>
</chain>
<chain id="120">
    <frequency>4</frequency>
    <phrase>ink
        <node>dry</node> on the
    </phrase>
</chain>
<chain id="121">
    <frequency>4</frequency>
    <phrase>is to
        <node>dry</node> it
    </phrase>
</chain>
<chain id="122">
    <frequency>4</frequency>
    <phrase>it a very
        <node>dry</node>
    </phrase>
</chain>
<chain id="123">
    <frequency>4</frequency>
    <phrase>just
        <node>dry</node> scientific facts
    </phrase>
</chain>
<chain id="124">
    <frequency>4</frequency>
    <phrase>people farming in
        <node>dry</node>
    </phrase>
</chain>
<chain id="125">
    <frequency>4</frequency>
    <phrase>rainfall and a
        <node>dry</node>
    </phrase>
</chain>
<chain id="126">
    <frequency>4</frequency>
    <phrase>ships in
        <node>dry</node> docks
    </phrase>
</chain>
<chain id="127">
    <frequency>4</frequency>
    <phrase>started to
        <node>dry</node> up
    </phrase>
</chain>
<chain id="128">
    <frequency>4</frequency>
    <phrase>the
        <node>dry</node> season has
    </phrase>
</chain>
<chain id="129">
    <frequency>4</frequency>
    <phrase>the
        <node>dry</node> season in
    </phrase>
</chain>
<chain id="130">
    <frequency>4</frequency>
    <phrase>to
```

```
            <node>dry</node> it and
        </phrase>
    </chain>
    <chain id="131">
        <frequency>4</frequency>
        <phrase>to
            <node>dry</node> out
        </phrase>
    </chain>
    <chain id="132">
        <frequency>4</frequency>
        <phrase>to
            <node>dry</node> out and
        </phrase>
    </chain>
    <chain id="133">
        <frequency>4</frequency>
        <phrase>to be tinder
            <node>dry</node>
        </phrase>
    </chain>
    <chain id="134">
        <frequency>4</frequency>
        <phrase>was the ink
            <node>dry</node>
        </phrase>
    </chain>
    <chain id="135">
        <frequency>3</frequency>
        <phrase>
            <node>dry</node> it and
        </phrase>
    </chain>
    <chain id="136">
        <frequency>3</frequency>
        <phrase>
            <node>dry</node> out and
        </phrase>
    </chain>
    <chain id="137">
        <frequency>3</frequency>
        <phrase>
            <node>dry</node> season in
        </phrase>
    </chain>
    <chain id="138">
        <frequency>3</frequency>
        <phrase>a very
            <node>dry</node>
        </phrase>
    </chain>
    <chain id="139">
        <frequency>3</frequency>
        <phrase>consecutive
            <node>dry</node> days
        </phrase>
    </chain>
    <chain id="140">
        <frequency>3</frequency>
        <phrase>during the
            <node>dry</node>
        </phrase>
    </chain>
    <chain id="141">
        <frequency>3</frequency>
        <phrase>fifteen consecutive
            <node>dry</node>
        </phrase>
    </chain>
    <chain id="142">
        <frequency>3</frequency>
        <phrase>ink
            <node>dry</node> on
        </phrase>
    </chain>
    <chain id="143">
        <frequency>3</frequency>
        <phrase>the coming
            <node>dry</node>
        </phrase>
    </chain>
    <chain id="144">
        <frequency>3</frequency>
        <phrase>the ink
            <node>dry</node>
        </phrase>
    </chain>
</chains>
<units_of_meaning>
    <uom id="0">
        <frequency>5</frequency>
        <phrase>would eat
            <node>dry</node> bread before
            <coll>begging</coll>
        </phrase>
    </uom>
</units_of_meaning>
<mwu_substitution>
    <substitution>
        <context frequencyband="8">areas</context>
        <entry frequency="10">dry</entry>
        <entry frequency="7">northern</entry>
        <entry frequency="7">tropical</entry>
        <entry frequency="6">Kurdish</entry>
        <entry frequency="6">four</entry>
        <entry frequency="6">troubled</entry>
        <entry frequency="5">Arab</entry>
        <entry frequency="5">contaminated</entry>
        <entry frequency="4">Muslim</entry>
        <entry frequency="4">both</entry>
        <entry frequency="4">coastal</entry>
        <entry frequency="4">designated</entry>
        <entry frequency="4">important</entry>
        <entry frequency="4">jungle</entry>
        <entry frequency="4">sensitive</entry>
        <entry frequency="4">strategic</entry>
        <entry frequency="4">vital</entry>
        <entry frequency="3">arid</entry>
        <entry frequency="3">built-up</entry>
        <entry frequency="3">country</entry>
        <entry frequency="3">crowded</entry>
        <entry frequency="3">deep</entry>
        <entry frequency="3">industrial</entry>
        <entry frequency="3">mixed</entry>
        <entry frequency="3">neighbouring</entry>
        <entry frequency="3">separate</entry>
        <entry frequency="3">slum</entry>
        <entry frequency="3">specific</entry>
        <entry frequency="3">to</entry>
        <entry frequency="3">uninhabited</entry>
        <entry frequency="3">wet</entry>
    </substitution>
    <substitution>
        <context frequencyband="5">out</context>
        <entry frequency="6">dry</entry>
        <entry frequency="3">bear</entry>
        <entry frequency="3">climb</entry>
        <entry frequency="3">drag</entry>
        <entry frequency="3">drive</entry>
        <entry frequency="3">eke</entry>
        <entry frequency="3">fill</entry>
        <entry frequency="3">leave</entry>
        <entry frequency="3">look</entry>
        <entry frequency="3">lose</entry>
        <entry frequency="3">pump</entry>
        <entry frequency="3">sit</entry>
        <entry frequency="3">test</entry>
        <entry frequency="3">throw</entry>
        <entry frequency="3">travel</entry>
        <entry frequency="3">working</entry>
    </substitution>
    <substitution>
        <context frequencyband="10">season</context>
        <entry frequency="48">dry</entry>
        <entry frequency="17">holiday</entry>
        <entry frequency="13">football</entry>
        <entry frequency="12">English</entry>
        <entry frequency="10">close</entry>
        <entry frequency="8">tourist</entry>
        <entry frequency="7">winter</entry>
        <entry frequency="5">Christmas</entry>
        <entry frequency="5">coming</entry>
        <entry frequency="5">festive</entry>
        <entry frequency="4">flood</entry>
        <entry frequency="3">growing</entry>
        <entry frequency="3">high</entry>
        <entry frequency="3">league</entry>
        <entry frequency="3">lean</entry>
        <entry frequency="3">monsoon</entry>
        <entry frequency="3">rainy</entry>
        <entry frequency="3">summer</entry>
        <entry frequency="3">wet</entry>
    </substitution>
</mwu_substitution>
<lexical_substitution>
    <arguments relation="VO" position="0">
```

```
<entry id="0">
    <score>11.258606162717703</score>
    <argument>bread</argument>
</entry>
<entry id="1">
    <score>10.174541897929227</score>
    <argument>wood</argument>
</entry>
<entry id="2">
    <score>9.793937895714258</score>
    <argument>season</argument>
</entry>
<entry id="3">
    <score>7.447135132187865</score>
    <argument>land</argument>
</entry>
<entry id="4">
    <score>7.339129867742813</score>
    <argument>areas</argument>
</entry>
<entry id="5">
    <score>6.319379584989493</score>
    <argument>area</argument>
</entry>
</arguments>
</lexical_substitution>
</dict_entry>
```

CHAPTER C

CASE STUDIES: XML OUTPUT

This appendix contains the data extracted for the case studies in chapter 6. Since the full output files are fairly large (see appendix B), they have been edited and all data not relevant to the case studies has been removed. This includes the usage pattern arguments; however, these are given in the substitution entries anyway.

C.1 Africa

```
<mwu_substitution>
    <substitution>
        <context frequencyband="7">BBC</context>
        <entry frequency="409">Africa</entry>
        <entry frequency="17">African</entry>
    </substitution>
    <substitution>
        <context frequencyband="6">East</context>
        <entry frequency="188">Africa</entry>
        <entry frequency="3">Asian</entry>
        <entry frequency="3">Berlin</entry>
        <entry frequency="3">Europe</entry>
        <entry frequency="3">Europeans</entry>
    </substitution>
    <substitution>
        <context frequencyband="8">North</context>
        <entry frequency="1325">Africa</entry>
        <entry frequency="23">African</entry>
        <entry frequency="19">America</entry>
    </substitution>
    <substitution>
        <context frequencyband="7">Our</context>
        <entry frequency="201">Africa</entry>
        <entry frequency="3">African</entry>
    </substitution>
    <substitution>
        <context frequencyband="7">South</context>
        <entry frequency="516">Africa</entry>
        <entry frequency="230">Korea</entry>
        <entry frequency="30">America</entry>
        <entry frequency="16">Africans</entry>
        <entry frequency="15">Asia</entry>
        <entry frequency="14">Wales</entry>
        <entry frequency="12">Korean</entry>
        <entry frequency="7">Georgia</entry>
        <entry frequency="5">African</entry>
        <entry frequency="5">Pacific</entry>
        <entry frequency="5">co-operation</entry>
        <entry frequency="4">Pole</entry>
        <entry frequency="4">and</entry>
        <entry frequency="3">Carolina</entry>
        <entry frequency="3">India</entry>
        <entry frequency="3">Koreans</entry>
        <entry frequency="3">Williamson</entry>
        <entry frequency="3">tend</entry>
    </substitution>
    <substitution>
        <context frequencyband="10">Southern</context>
        <entry frequency="1532">Africa</entry>
        <entry frequency="18">African</entry>
        <entry frequency="3">AFrica</entry>
    </substitution>
    <substitution>
        <context frequencyband="7">West</context>
        <entry frequency="444">Africa</entry>
        <entry frequency="5">African</entry>
        <entry frequency="3">Europe</entry>
    </substitution>
</mwu_substitution>
<lexical_substitution>
    <substitution relation="NN" position="0">
        <entry id="0">
            <score>1.000999000999001</score>
            <substitute>africa</substitute>
            <arguments>madagascar rehabilitation watch
                mike whites homelands hunger inkatha
                text exiles townships fm f15 wooldridge
                keane resignations isolation hiett
                back emergency catholics athletes violence
                blane churches return olympics blacks
                south blunt township participation
                while law s deaths communities</arguments>
        </entry>
        <entry id="1">
            <score>0.2701519462968 54</score>
            <substitute>south</substitute>
            <arguments>exiles homelands fm resignations
                back isolation athletes violence churches
```

```
            olympics</arguments>
        </entry>
        <entry id="2">
            <score>0.1350820159161768</score>
            <substitute>black</substitute>
            <arguments>township townships homelands
                south communities</arguments>
        </entry>
        <entry id="3">
            <score>0.1351027612370625</score>
            <substitute>correspondent</substitute>
            <arguments>mike blane keane wooldridge
                blunt</arguments>
        </entry>
        <entry id="4">
            <score>0.1351379772432404</score>
            <substitute>india</substitute>
            <arguments>resignations communities violence
                s deaths</arguments>
        </entry>
        <entry id="5">
            <score>0.10804699925765618</score>
            <substitute>many</substitute>
            <arguments>whites churches blacks catholics</arguments>
        </entry>
        <entry id="6">
            <score>0.10811759538571944</score>
            <substitute>s</substitute>
            <arguments>emergency churches back return</arguments>
        </entry>
        <entry id="7">
            <score>0.08106628681538129</score>
            <substitute>all</substitute>
            <arguments>blacks exiles communities</arguments>
        </entry>
        <entry id="8">
            <score>0.08103531596354563</score>
            <substitute>bush</substitute>
            <arguments>emergency s south</arguments>
        </entry>
        <entry id="9">
            <score>0.08101877252619585</score>
            <substitute>country</substitute>
            <arguments>townships isolation communities</arguments>
        </entry>
        <entry id="10">
            <score>0.0810417714040407</score>
            <substitute>iraq</substitute>
            <arguments>participation back isolation</arguments>
        </entry>
        <entry id="11">
            <score>0.08106970649691088</score>
            <substitute>talks</substitute>
            <arguments>text keane south</arguments>
        </entry>
    </substitution>
    <substitution relation="SV" position="0">
        <entry id="0">
            <score>1.000999000999001</score>
            <substitute>africa</substitute>
            <arguments>re-admitted readmitted demoted
                watch embarked regained expelled package
                granted denounced pressing facing aid
                announced needs needed allowed</arguments>
        </entry>
        <entry id="1">
            <score>0.1761183801778023</score>
            <substitute>being</substitute>
            <arguments>expelled granted allowed</arguments>
        </entry>
        <entry id="2">
            <score>0.176161150585703</score>
            <substitute>france</substitute>
            <arguments>expelled granted pressing</arguments>
        </entry>
        <entry id="3">
            <score>0.1760815881464141</score>
            <substitute>party</substitute>
            <arguments>embarked denounced facing</arguments>
        </entry>
    </substitution>
    <substitution relation="VO" position="1">
        <entry id="0">
            <score>1.000999000999001</score>
            <substitute>africa</substitute>
            <arguments>post-apartheid re-admitted sub-saharan
                re-admit readmit betray toured evicted
                isolate covered visited fled visit
                black hit wanted beat help bring left</arguments>
        </entry>
        <entry id="1">
            <score>0.25006398943681174</score>
            <substitute>area</substitute>
            <arguments>fled visited hit covered visit</arguments>
        </entry>
        <entry id="2">
            <score>0.2501978973407545</score>
            <substitute>china</substitute>
            <arguments>visit isolate visited fled beat</arguments>
        </entry>
        <entry id="3">
            <score>0.20042338613767183</score>
            <substitute>baghdad</substitute>
            <arguments>isolate visited visit left</arguments>
        </entry>
        <entry id="4">
            <score>0.20010888582317155</score>
            <substitute>britain</substitute>
            <arguments>isolate visit left hit</arguments>
        </entry>
        <entry id="5">
            <score>0.2005375576804148</score>
            <substitute>egypt</substitute>
            <arguments>visit visited beat left</arguments>
        </entry>
        <entry id="6">
            <score>0.20004834101472757</score>
            <substitute>north</substitute>
            <arguments>fled visited hit wanted</arguments>
        </entry>
        <entry id="7">
            <score>0.20027115741401455</score>
            <substitute>pakistan</substitute>
            <arguments>hit visit beat left</arguments>
        </entry>
        <entry id="8">
            <score>0.2001902859045716</score>
            <substitute>region</substitute>
            <arguments>covered visited visit hit</arguments>
        </entry>
        <entry id="9">
            <score>0.20023483010495996</score>
            <substitute>romania</substitute>
            <arguments>beat visit left bring</arguments>
        </entry>
        <entry id="10">
            <score>0.1500963322391894</score>
            <substitute>arabia</substitute>
            <arguments>visited visit left</arguments>
        </entry>
        <entry id="11">
            <score>0.14996402139259282</score>
            <substitute>areas</substitute>
            <arguments>hit visited visit</arguments>
        </entry>
        <entry id="12">
            <score>0.15022913594342166</score>
            <substitute>argentina</substitute>
            <arguments>beat visit left</arguments>
        </entry>
        <entry id="13">
            <score>0.15002017951326707</score>
            <substitute>capital</substitute>
            <arguments>hit left visited</arguments>
        </entry>
        <entry id="14">
            <score>0.14999511766429058</score>
            <substitute>country</substitute>
            <arguments>fled toured visit</arguments>
        </entry>
        <entry id="15">
            <score>0.15033220747506462</score>
            <substitute>czechoslovakia</substitute>
            <arguments>beat help visit</arguments>
        </entry>
        <entry id="16">
            <score>0.15014192157049303</score>
            <substitute>east</substitute>
            <arguments>visit black bring</arguments>
        </entry>
        <entry id="17">
            <score>0.15009276437847865</score>
            <substitute>india</substitute>
            <arguments>visited help left</arguments>
        </entry>
        <entry id="18">
```

```
            <score>0.1501126579048657</score>
            <substitute>iran</substitute>
            <arguments>visit hit visited</arguments>
        </entry>
        <entry id="19">
            <score>0.15004972878727033</score>
            <substitute>iraq</substitute>
            <arguments>isolate visit fled</arguments>
        </entry>
        <entry id="20">
            <score>0.15027908599337172</score>
            <substitute>jordan</substitute>
            <arguments>visit left help</arguments>
        </entry>
        <entry id="21">
            <score>0.15027987885130742</score>
            <substitute>korea</substitute>
            <arguments>beat visited visit</arguments>
        </entry>
        <entry id="22">
            <score>0.15019396188227357</score>
            <substitute>london</substitute>
            <arguments>visited left visit</arguments>
        </entry>
        <entry id="23">
            <score>0.15011079146417494</score>
            <substitute>moscow</substitute>
            <arguments>visit left visited</arguments>
        </entry>
        <entry id="24">
            <score>0.1502393710185918</score>
            <substitute>peking</substitute>
            <arguments>visit visited left</arguments>
        </entry>
        <entry id="25">
            <score>0.15013255143125273</score>
            <substitute>poland</substitute>
            <arguments>visited visit help</arguments>
        </entry>
        <entry id="26">
            <score>0.15019742162599306</score>
            <substitute>south</substitute>
            <arguments>black visit beat</arguments>
        </entry>
        <entry id="27">
            <score>0.15005290276226238</score>
            <substitute>town</substitute>
            <arguments>fled covered visited</arguments>
        </entry>
        <entry id="28">
            <score>0.15015204575644137</score>
            <substitute>washington</substitute>
            <arguments>visit visited left</arguments>
        </entry>
    </substitution>
</lexical_substitution>
```

C.2 Germany

```
<mwu_substitution>
    <substitution>
        <context frequencyband="6">East</context>
        <entry frequency="616">Germany</entry>
        <entry frequency="88">Berlin</entry>
        <entry frequency="29">Germans</entry>
        <entry frequency="25">Asia</entry>
        <entry frequency="18">Jerusalem</entry>
        <entry frequency="16">Europe</entry>
        <entry frequency="11">Beirut</entry>
        <entry frequency="11">German</entry>
        <entry frequency="10">Asian</entry>
        <entry frequency="10">Timor</entry>
        <entry frequency="7">Africa</entry>
        <entry frequency="6">Anglia</entry>
        <entry frequency="6">European</entry>
        <entry frequency="6">oilfields</entry>
        <entry frequency="6">peace</entry>
        <entry frequency="4">politics</entry>
        <entry frequency="3">Europeans</entry>
    </substitution>
    <substitution>
        <context frequencyband="6">Union</context>
        <entry frequency="5">Germany</entry>
        <entry frequency="4">Britain</entry>
        <entry frequency="3">India</entry>
        <entry frequency="3">Vietnam</entry>
    </substitution>
    <substitution>
        <context frequencyband="7">West</context>
        <entry frequency="946">Germany</entry>
        <entry frequency="220">Bank</entry>
        <entry frequency="51">Germans</entry>
        <entry frequency="42">German</entry>
        <entry frequency="30">Berlin</entry>
        <entry frequency="28">Indies</entry>
        <entry frequency="25">relations</entry>
        <entry frequency="19">Africa</entry>
        <entry frequency="12">and</entry>
        <entry frequency="8">Beirut</entry>
        <entry frequency="8">summit</entry>
        <entry frequency="7">End</entry>
        <entry frequency="7">had</entry>
        <entry frequency="5">England</entry>
        <entry frequency="5">Yorkshire</entry>
        <entry frequency="5">talks</entry>
        <entry frequency="3">continues</entry>
        <entry frequency="3">must</entry>
        <entry frequency="3">not</entry>
        <entry frequency="3">now</entry>
        <entry frequency="3">wanted</entry>
    </substitution>
    <substitution>
        <context frequencyband="6">called</context>
        <entry frequency="3">Belgium</entry>
        <entry frequency="3">Germany</entry>
        <entry frequency="3">Japan</entry>
        <entry frequency="3">Muslims</entry>
        <entry frequency="3">NATO</entry>
        <entry frequency="3">deputies</entry>
        <entry frequency="3">miners</entry>
        <entry frequency="3">residents</entry>
        <entry frequency="3">society</entry>
        <entry frequency="3">whites</entry>
    </substitution>
    <substitution>
        <context frequencyband="2">in</context>
        <entry frequency="3">Britain</entry>
        <entry frequency="3">France</entry>
        <entry frequency="3">Germany</entry>
        <entry frequency="3">competition</entry>
    </substitution>
    <substitution>
        <context frequencyband="6">should</context>
        <entry frequency="21">Germany</entry>
        <entry frequency="9">people</entry>
        <entry frequency="6">supplies</entry>
        <entry frequency="6">that</entry>
        <entry frequency="5">children</entry>
        <entry frequency="5">force</entry>
        <entry frequency="5">she</entry>
        <entry frequency="4">Kashmir</entry>
        <entry frequency="4">action</entry>
        <entry frequency="4">drugs</entry>
        <entry frequency="4">elections</entry>
        <entry frequency="4">talks</entry>
        <entry frequency="4">these</entry>
        <entry frequency="3">Kuwait</entry>
        <entry frequency="3">Pretoria</entry>
        <entry frequency="3">food</entry>
        <entry frequency="3">he</entry>
        <entry frequency="3">mission</entry>
        <entry frequency="3">parliament</entry>
        <entry frequency="3">pressure</entry>
        <entry frequency="3">priority</entry>
        <entry frequency="3">service</entry>
        <entry frequency="3">stars</entry>
        <entry frequency="3">their</entry>
```

```
        <entry frequency="3">who</entry>
    </substitution>
    <substitution>
        <context frequencyband="9">united</context>
        <entry frequency="639">Germany</entry>
        <entry frequency="25">front</entry>
        <entry frequency="4">campaign</entry>
        <entry frequency="3">approach</entry>
    </substitution>
</mwu_substitution>
<lexical_substitution>
    <substitution relation="AN" position="1">
        <entry id="0">
            <score>1.000999000999001</score>
            <substitute>germany</substitute>
            <arguments>reunified reunited uniting unified
                nazi present-day postwar neutral post-war
                sovereign then divided capitalist powerful</arguments>
        </entry>
        <entry id="1">
            <score>0.286191489199008</score>
            <substitute>country</substitute>
            <arguments>divided capitalist neutral unified</arguments>
        </entry>
        <entry id="2">
            <score>0.2140919122285582</score>
            <substitute>body</substitute>
            <arguments>sovereign neutral powerful</arguments>
        </entry>
        <entry id="3">
            <score>0.2141666456792507</score>
            <substitute>nation</substitute>
            <arguments>divided unified powerful</arguments>
        </entry>
        <entry id="4">
            <score>0.21416643751138525</score>
            <substitute>period</substitute>
            <arguments>postwar post-war nazi</arguments>
        </entry>
    </substitution>
    <substitution relation="NN" position="0">
        <entry id="0">
            <score>1.000999000999001</score>
            <substitute>germany</substitute>
            <arguments>langer krabbe ludwig bundesbank
                cueto accession membership wouldn't
                heartland stich electricity athletes
                unification unity chancellor integration
                status past stability partners means
                prosecutor division neighbours freedom
                right borders democrats democracy being
                genscher border opposition constitution
                relationship coalition currency image
                embassy court success network history</arguments>
        </entry>
        <entry id="1">
            <score>0.16317669149679334</score>
            <substitute>german</substitute>
            <arguments>unification bundesbank unity
                chancellor membership currency embassy</arguments>
        </entry>
        <entry id="2">
            <score>0.16326805803367733</score>
            <substitute>poland</substitute>
            <arguments>integration democracy borders
                stability border history constitution</arguments>
        </entry>
        <entry id="3">
            <score>0.14007934928409801</score>
            <substitute>japan</substitute>
            <arguments>constitution past neighbours
                success relationship image</arguments>
        </entry>
        <entry id="4">
            <score>0.11603153654503048</score>
            <substitute>country</substitute>
            <arguments>electricity stability past unity
                democracy</arguments>
        </entry>
        <entry id="5">
            <score>0.11603674003508996</score>
            <substitute>future</substitute>
            <arguments>stability status relationship
                membership constitution</arguments>
        </entry>
        <entry id="6">
            <score>0.11607621330292639</score>
            <substitute>republic</substitute>
            <arguments>right constitution borders membership
                status</arguments>
        </entry>
        <entry id="7">
            <score>0.09304787450214269</score>
            <substitute>israel</substitute>
            <arguments>right borders neighbours membership</arguments>
        </entry>
        <entry id="8">
            <score>0.09319518697350168</score>
            <substitute>liberal</substitute>
            <arguments>democrats democracy opposition
                constitution</arguments>
        </entry>
        <entry id="9">
            <score>0.09313769573032953</score>
            <substitute>lithuania</substitute>
            <arguments>right status border neighbours</arguments>
        </entry>
        <entry id="10">
            <score>0.09310831790445201</score>
            <substitute>pakistan</substitute>
            <arguments>border borders democracy court</arguments>
        </entry>
        <entry id="11">
            <score>0.0701876782588137</score>
            <substitute>belgian</substitute>
            <arguments>prosecutor embassy border</arguments>
        </entry>
        <entry id="12">
            <score>0.07004782621077348</score>
            <substitute>britain</substitute>
            <arguments>electricity chancellor membership</arguments>
        </entry>
        <entry id="13">
            <score>0.07012401147891542</score>
            <substitute>canadian</substitute>
            <arguments>constitution unity embassy</arguments>
        </entry>
        <entry id="14">
            <score>0.07001773203350084</score>
            <substitute>china</substitute>
            <arguments>stability status image</arguments>
        </entry>
        <entry id="15">
            <score>0.07004579005855484</score>
            <substitute>community</substitute>
            <arguments>partners membership currency</arguments>
        </entry>
        <entry id="16">
            <score>0.0702225892500734</score>
            <substitute>democrat</substitute>
            <arguments>chancellor partners opposition</arguments>
        </entry>
        <entry id="17">
            <score>0.07007715932512067</score>
            <substitute>east</substitute>
            <arguments>ludwig accession electricity</arguments>
        </entry>
        <entry id="18">
            <score>0.07010332866569653</score>
            <substitute>egyptian</substitute>
            <arguments>border embassy court</arguments>
        </entry>
        <entry id="19">
            <score>0.07005668188681334</score>
            <substitute>european</substitute>
            <arguments>integration currency unity</arguments>
        </entry>
        <entry id="20">
            <score>0.07017563669051716</score>
            <substitute>hungary</substitute>
            <arguments>democracy membership border</arguments>
        </entry>
        <entry id="21">
            <score>0.07006141950267591</score>
            <substitute>japanese</substitute>
            <arguments>currency embassy constitution</arguments>
        </entry>
        <entry id="22">
            <score>0.07033080294702393</score>
            <substitute>kohl</substitute>
            <arguments>democrats coalition partners</arguments>
        </entry>
        <entry id="23">
            <score>0.07000875898489912</score>
            <substitute>kong</substitute>
            <arguments>constitution image success</arguments>
```

```
        </entry>
        <entry id="24">
            <score>0.07000649884605883</score>
            <substitute>party</substitute>
            <arguments>democracy unity membership</arguments>
        </entry>
        <entry id="25">
            <score>0.07015943886979828</score>
            <substitute>post</substitute>
            <arguments>borders constitution border</arguments>
        </entry>
        <entry id="26">
            <score>0.07014559989492035</score>
            <substitute>serbia</substitute>
            <arguments>borders border constitution</arguments>
        </entry>
        <entry id="27">
            <score>0.07010492779142734</score>
            <substitute>yugoslavia</substitute>
            <arguments>borders neighbours constitution</arguments>
        </entry>
    </substitution>
    <substitution relation="SV" position="0">
        <entry id="0">
            <score>1.000999000999001</score>
            <substitute>germany</substitute>
            <arguments>anchored accede sulking belong
                unite leading prepares join pressing
                fit nato spoken stay voting recognised
                giving part united stepped remain pledged
                banned signed joined play send sending</arguments>
        </entry>
        <entry id="1">
            <score>0.18507284809805818</score>
            <substitute>all</substitute>
            <arguments>unite nato belong fit part</arguments>
        </entry>
        <entry id="2">
            <score>0.18512106534084558</score>
            <substitute>britain</substitute>
            <arguments>join sending joined play send</arguments>
        </entry>
        <entry id="3">
            <score>0.18512546300002747</score>
            <substitute>countries</substitute>
            <arguments>signed join pledged recognised
                pressing</arguments>
        </entry>
        <entry id="4">
            <score>0.1480900458922437</score>
            <substitute>members</substitute>
            <arguments>voting belong pressing join</arguments>
        </entry>
        <entry id="5">
            <score>0.14811106285760664</score>
            <substitute>nations</substitute>
            <arguments>united sending send play</arguments>
        </entry>
        <entry id="6">
            <score>0.1481008598355537</score>
            <substitute>party</substitute>
            <arguments>unite prepares leading banned</arguments>
        </entry>
        <entry id="7">
            <score>0.14805947428548707</score>
            <substitute>will</substitute>
            <arguments>play join remain stay</arguments>
        </entry>
        <entry id="8">
            <score>0.1110458310611099</score>
            <substitute>community</substitute>
            <arguments>send recognised sending</arguments>
        </entry>
        <entry id="9">
            <score>0.11095899639621566</score>
            <substitute>forces</substitute>
            <arguments>stepped stay remain</arguments>
        </entry>
        <entry id="10">
            <score>0.11114851332242635</score>
            <substitute>france</substitute>
            <arguments>sending pressing joined</arguments>
        </entry>
        <entry id="11">
            <score>0.1110118948280786 5</score>
            <substitute>troops</substitute>
            <arguments>stay remain join</arguments>
        </entry>
    </substitution>
    <substitution relation="VO" position="1">
        <entry id="0">
            <score>1.000999000999001</score>
            <substitute>germany</substitute>
            <arguments>consigned administered bind
                absorb looked expects united fear allows
                wants accept represented defeated meeting
                occupied visiting divided lose want</arguments>
        </entry>
        <entry id="1">
            <score>0.2110209628428252</score>
            <substitute>states</substitute>
            <arguments>united visiting allows represented</arguments>
        </entry>
        <entry id="2">
            <score>0.1580174439367877 4</score>
            <substitute>return</substitute>
            <arguments>wants want accept</arguments>
        </entry>
        <entry id="3">
            <score>0.15798238281589516</score>
            <substitute>war</substitute>
            <arguments>fear want lose</arguments>
        </entry>
    </substitution>
</lexical_substitution>
```

C.3 Monday

```
<mwu_substitution>
    <substitution>
        <context frequencyband="12">ARABIC</context>
        <entry frequency="6">Monday</entry>
        <entry frequency="6">Thursday</entry>
        <entry frequency="6">Tuesday</entry>
        <entry frequency="5">Friday</entry>
    </substitution>
    <substitution>
        <context frequencyband="13">Mitford</context>
        <entry frequency="6">Friday</entry>
        <entry frequency="6">Monday</entry>
    </substitution>
    <substitution>
        <context frequencyband="8">On</context>
        <entry frequency="35">Monday</entry>
        <entry frequency="23">Saturday</entry>
        <entry frequency="21">Friday</entry>
        <entry frequency="15">Sunday</entry>
        <entry frequency="15">Wednesday</entry>
        <entry frequency="13">Thursday</entry>
        <entry frequency="13">Tuesday</entry>
        <entry frequency="7">May</entry>
        <entry frequency="5">October</entry>
        <entry frequency="5">entering</entry>
        <entry frequency="4">July</entry>
        <entry frequency="4">paper</entry>
        <entry frequency="4">to</entry>
        <entry frequency="3">January</entry>
        <entry frequency="3">November</entry>
        <entry frequency="3">all</entry>
        <entry frequency="3">board</entry>
    </substitution>
    <substitution>
        <context frequencyband="8">Peking</context>
        <entry frequency="12">Monday</entry>
        <entry frequency="3">Tuesday</entry>
    </substitution>
```

```
<substitution>
    <context frequencyband="6">after</context>
    <entry frequency="5">Monday</entry>
    <entry frequency="5">Sunday</entry>
    <entry frequency="5">bail</entry>
    <entry frequency="4">Wednesday</entry>
    <entry frequency="3">course</entry>
    <entry frequency="3">him</entry>
    <entry frequency="3">long</entry>
    <entry frequency="3">penalties</entry>
</substitution>
<substitution>
    <context frequencyband="10">afternoon</context>
    <entry frequency="16">Monday</entry>
    <entry frequency="15">Saturday</entry>
    <entry frequency="15">Tuesday</entry>
    <entry frequency="12">Thursday</entry>
    <entry frequency="8">the</entry>
    <entry frequency="7">Wednesday</entry>
</substitution>
<substitution>
    <context frequencyband="7">announced</context>
    <entry frequency="6">Monday</entry>
    <entry frequency="6">Wednesday</entry>
    <entry frequency="5">Friday</entry>
    <entry frequency="4">Thursday</entry>
</substitution>
<substitution>
    <context frequencyband="4">at</context>
    <entry frequency="4">Monday</entry>
    <entry frequency="4">duty</entry>
    <entry frequency="4">show</entry>
    <entry frequency="3">display</entry>
    <entry frequency="3">view</entry>
</substitution>
<substitution>
    <context frequencyband="6">between</context>
    <entry frequency="6">Monday</entry>
    <entry frequency="6">trade</entry>
    <entry frequency="5">Saturday</entry>
    <entry frequency="4">Thursday</entry>
    <entry frequency="4">Tuesday</entry>
    <entry frequency="4">Wednesday</entry>
    <entry frequency="3">Sunday</entry>
    <entry frequency="3">co-operation</entry>
</substitution>
<substitution>
    <context frequencyband="5">but</context>
    <entry frequency="4">Monday</entry>
    <entry frequency="4">Tuesday</entry>
    <entry frequency="4">all</entry>
    <entry frequency="3">Saturday</entry>
    <entry frequency="3">this</entry>
</substitution>
<substitution>
    <context frequencyband="8">discuss</context>
    <entry frequency="31">Monday</entry>
    <entry frequency="12">Tuesday</entry>
    <entry frequency="11">Thursday</entry>
    <entry frequency="9">Sunday</entry>
    <entry frequency="8">Wednesday</entry>
    <entry frequency="5">Saturday</entry>
</substitution>
<substitution>
    <context frequencyband="10">evening</context>
    <entry frequency="32">Monday</entry>
    <entry frequency="30">Wednesday</entry>
    <entry frequency="28">Sunday</entry>
    <entry frequency="27">Thursday</entry>
    <entry frequency="24">Friday</entry>
    <entry frequency="24">Saturday</entry>
    <entry frequency="9">the</entry>
    <entry frequency="5">an</entry>
    <entry frequency="4">this</entry>
</substitution>
<substitution>
    <context frequencyband="8">hours</context>
    <entry frequency="30">Monday</entry>
    <entry frequency="24">Thursday</entry>
    <entry frequency="24">Tuesday</entry>
    <entry frequency="21">Saturday</entry>
    <entry frequency="13">Friday</entry>
    <entry frequency="12">Sunday</entry>
    <entry frequency="12">tomorrow</entry>
</substitution>
<substitution>
    <context frequencyband="9">killing</context>
    <entry frequency="7">Monday</entry>
    <entry frequency="4">Sunday</entry>
    <entry frequency="3">Wednesday</entry>
</substitution>
<substitution>
    <context frequencyband="9">late</context>
    <entry frequency="14">Monday</entry>
    <entry frequency="5">Sunday</entry>
    <entry frequency="4">Saturday</entry>
    <entry frequency="4">Thursday</entry>
    <entry frequency="4">Tuesday</entry>
    <entry frequency="3">Friday</entry>
</substitution>
<substitution>
    <context frequencyband="8">morning</context>
    <entry frequency="62">Monday</entry>
    <entry frequency="49">Friday</entry>
    <entry frequency="47">Tuesday</entry>
    <entry frequency="44">Sunday</entry>
    <entry frequency="40">Thursday</entry>
    <entry frequency="39">Wednesday</entry>
    <entry frequency="34">Saturday</entry>
    <entry frequency="19">the</entry>
    <entry frequency="5">tomorrow</entry>
    <entry frequency="4">early</entry>
    <entry frequency="3">a</entry>
</substitution>
<substitution>
    <context frequencyband="8">night</context>
    <entry frequency="64">Monday</entry>
    <entry frequency="63">Friday</entry>
    <entry frequency="60">Saturday</entry>
    <entry frequency="52">Thursday</entry>
    <entry frequency="8">Census</entry>
    <entry frequency="4">a</entry>
    <entry frequency="4">all</entry>
    <entry frequency="3">census</entry>
    <entry frequency="3">that</entry>
</substitution>
<substitution>
    <context frequencyband="3">on</context>
    <entry frequency="37">Monday</entry>
    <entry frequency="25">Friday</entry>
    <entry frequency="20">Wednesday</entry>
    <entry frequency="14">Tuesday</entry>
    <entry frequency="13">Sunday</entry>
    <entry frequency="12">him</entry>
    <entry frequency="10">Sundays</entry>
    <entry frequency="10">Thursday</entry>
    <entry frequency="10">education</entry>
    <entry frequency="9">television</entry>
    <entry frequency="9">them</entry>
    <entry frequency="8">fire</entry>
    <entry frequency="8">it</entry>
    <entry frequency="8">radio</entry>
    <entry frequency="8">sale</entry>
    <entry frequency="8">strike</entry>
    <entry frequency="8">trade</entry>
    <entry frequency="7">Saturday</entry>
    <entry frequency="7">completion</entry>
    <entry frequency="7">condition</entry>
    <entry frequency="7">me</entry>
    <entry frequency="7">most</entry>
    <entry frequency="7">reports</entry>
    <entry frequency="7">that</entry>
    <entry frequency="6">hopes</entry>
    <entry frequency="6">many</entry>
    <entry frequency="6">offer</entry>
    <entry frequency="6">sales</entry>
    <entry frequency="6">stage</entry>
    <entry frequency="5">any</entry>
    <entry frequency="5">grounds</entry>
    <entry frequency="5">imports</entry>
    <entry frequency="5">members</entry>
    <entry frequency="5">political</entry>
    <entry frequency="5">time</entry>
    <entry frequency="5">two</entry>
    <entry frequency="4">Filipinos</entry>
    <entry frequency="4">Iraq</entry>
    <entry frequency="4">all</entry>
    <entry frequency="4">changes</entry>
    <entry frequency="4">crime</entry>
    <entry frequency="4">details</entry>
    <entry frequency="4">events</entry>
    <entry frequency="4">foot</entry>
    <entry frequency="4">himself</entry>
    <entry frequency="4">parliament</entry>
    <entry frequency="4">receipt</entry>
    <entry frequency="4">research</entry>
```

```
            <entry frequency="4">transfer</entry>
            <entry frequency="4">trial</entry>
            <entry frequency="3">Britain</entry>
            <entry frequency="3">Europe</entry>
            <entry frequency="3">SAT-M</entry>
            <entry frequency="3">TV</entry>
            <entry frequency="3">Washington</entry>
            <entry frequency="3">abortion</entry>
            <entry frequency="3">access</entry>
            <entry frequency="3">anything</entry>
            <entry frequency="3">barnacles</entry>
            <entry frequency="3">board</entry>
            <entry frequency="3">communication</entry>
            <entry frequency="3">display</entry>
            <entry frequency="3">exports</entry>
            <entry frequency="3">finding</entry>
            <entry frequency="3">hand</entry>
            <entry frequency="3">hands</entry>
            <entry frequency="3">hearing</entry>
            <entry frequency="3">income</entry>
            <entry frequency="3">issues</entry>
            <entry frequency="3">location</entry>
            <entry frequency="3">occasion</entry>
            <entry frequency="3">on</entry>
            <entry frequency="3">patrol</entry>
            <entry frequency="3">people</entry>
            <entry frequency="3">questions</entry>
            <entry frequency="3">reading</entry>
            <entry frequency="3">record</entry>
            <entry frequency="3">something</entry>
            <entry frequency="3">tourism</entry>
            <entry frequency="3">transport</entry>
            <entry frequency="3">us</entry>
            <entry frequency="3">view</entry>
            <entry frequency="3">voters</entry>
            <entry frequency="3">well</entry>
        </substitution>
        <substitution>
            <context frequencyband="6">when</context>
            <entry frequency="18">Monday</entry>
            <entry frequency="14">Sunday</entry>
            <entry frequency="12">Saturday</entry>
            <entry frequency="12">Wednesday</entry>
            <entry frequency="11">Friday</entry>
            <entry frequency="7">Tuesday</entry>
            <entry frequency="4">them</entry>
            <entry frequency="3">Thursday</entry>
        </substitution>
    </mwu_substitution>
    <lexical_substitution>
        <substitution relation="NN" position="0">
            <entry id="0">
                <score>1.000999000999001</score>
                <substitute>monday</substitute>
                <arguments>dec jul morning evening afternoon
                    night killings shooting clashes shootings
                    killing concert rebellion earthquake
                    october coup march june september may
                    violence august newspapers events incident
                    disaster papers</arguments>
            </entry>
            <entry id="1">
                <score>0.44462107093686043</score>
                <substitute>thursday</substitute>
                <arguments>dec jul evening morning afternoon
                    killing september june october earthquake
                    may incident</arguments>
            </entry>
            <entry id="2">
                <score>0.4442405964962356</score>
                <substitute>week</substitute>
                <arguments>coup shootings events earthquake
                    shooting rebellion killings killing
                    violence clashes incident disaster</arguments>
            </entry>
            <entry id="3">
                <score>0.3704464847321 99</score>
                <substitute>friday</substitute>
                <arguments>afternoon evening morning night
                    clashes september october may june
                    violence</arguments>
            </entry>
            <entry id="4">
                <score>0.37028844171701314</score>
                <substitute>sunday</substitute>
                <arguments>afternoon morning evening night
                    killing papers newspapers killings
                    june coup</arguments>
            </entry>
            <entry id="5">
                <score>0.3706556048993024</score>
                <substitute>tuesday</substitute>
                <arguments>jul evening morning afternoon
                    may earthquake october september june
                    papers</arguments>
            </entry>
            <entry id="6">
                <score>0.33362260988311404</score>
                <substitute>wednesday</substitute>
                <arguments>morning earthquake evening night
                    afternoon may august clashes october</arguments>
            </entry>
            <entry id="7">
                <score>0.22239276440957112</score>
                <substitute>saturday</substitute>
                <arguments>afternoon morning night evening
                    violence coup</arguments>
            </entry>
            <entry id="8">
                <score>0.18544788544788543</score>
                <substitute>arabic</substitute>
                <arguments>october june september may august</arguments>
            </entry>
            <entry id="9">
                <score>0.1851613994471137</score>
                <substitute>night</substitute>
                <arguments>shooting earthquake violence
                    incident events</arguments>
            </entry>
            <entry id="10">
                <score>0.18511245839181062</score>
                <substitute>service</substitute>
                <arguments>october june september august
                    may</arguments>
            </entry>
            <entry id="11">
                <score>0.18503946750699996</score>
                <substitute>st</substitute>
                <arguments>march august october september
                    june</arguments>
            </entry>
            <entry id="12">
                <score>0.18503718503718505</score>
                <substitute>tim</substitute>
                <arguments>september october may august
                    june</arguments>
            </entry>
            <entry id="13">
                <score>0.14813390452488195</score>
                <substitute>jerusalem</substitute>
                <arguments>killings shootings incident
                    violence</arguments>
            </entry>
            <entry id="14">
                <score>0.14822615424119184</score>
                <substitute>morning</substitute>
                <arguments>newspapers papers incident events</arguments>
            </entry>
            <entry id="15">
                <score>0.1480669202313755</score>
                <substitute>one</substitute>
                <arguments>incident evening afternoon morning</arguments>
            </entry>
            <entry id="16">
                <score>0.14821998588232355</score>
                <substitute>rd</substitute>
                <arguments>october june august may</arguments>
            </entry>
            <entry id="17">
                <score>0.1480091163252829</score>
                <substitute>year</substitute>
                <arguments>rebellion events june killings</arguments>
            </entry>
            <entry id="18">
                <score>0.1110317236690863</score>
                <substitute>football</substitute>
                <arguments>violence disaster june</arguments>
            </entry>
            <entry id="19">
                <score>0.1111078583606056</score>
                <substitute>month</substitute>
                <arguments>earthquake coup violence</arguments>
            </entry>
            <entry id="20">
                <score>0.11118669690098261</score>
                <substitute>th</substitute>
```

```
            <arguments>september august october</arguments>
        </entry>
        <entry id="21">
            <score>0.11113715529299945</score>
            <substitute>weekend</substitute>
            <arguments>violence killings events</arguments>
        </entry>
    </substitution>
    <substitution relation="NN" position="1">
        <entry id="0">
            <score>1.000999000999001</score>
            <substitute>monday</substitute>
            <arguments>menu collection unit holiday
                service</arguments>
        </entry>
        <entry id="1">
            <score>0.6002283430854859</score>
            <substitute>wednesday</substitute>
            <arguments>menu unit service</arguments>
        </entry>
    </substitution>
</lexical_substitution>
```

C.4 Smith

```
<mwu_substitution>
    <substitution>
        <context frequencyband="8">Peter</context>
        <entry frequency="36">Smith</entry>
        <entry frequency="12">Laverock</entry>
        <entry frequency="8">Murray</entry>
        <entry frequency="7">Pockley</entry>
        <entry frequency="6">Goodwin</entry>
        <entry frequency="5">Hunt</entry>
        <entry frequency="3">Bowes</entry>
        <entry frequency="3">Nares</entry>
    </substitution>
    <substitution>
        <context frequencyband="4">said</context>
        <entry frequency="6">Alatas</entry>
        <entry frequency="6">Cook</entry>
        <entry frequency="6">Gummer</entry>
        <entry frequency="6">Heath</entry>
        <entry frequency="6">Lacayo</entry>
        <entry frequency="6">Smith</entry>
        <entry frequency="6">Walesa</entry>
        <entry frequency="5">Arafat</entry>
        <entry frequency="5">Arens</entry>
        <entry frequency="5">Aziz</entry>
        <entry frequency="5">Bessmertnykh</entry>
        <entry frequency="5">Genscher</entry>
        <entry frequency="5">Hattersley</entry>
        <entry frequency="5">Heseltine</entry>
        <entry frequency="5">Iliescu</entry>
        <entry frequency="5">MacGregor</entry>
        <entry frequency="5">Mazowiecki</entry>
        <entry frequency="5">Patten</entry>
        <entry frequency="4">Bhattarai</entry>
        <entry frequency="4">Botha</entry>
        <entry frequency="4">Brooke</entry>
        <entry frequency="4">Commons</entry>
        <entry frequency="4">Geremek</entry>
        <entry frequency="4">Gray</entry>
        <entry frequency="4">Hawke</entry>
        <entry frequency="4">Hrawi</entry>
        <entry frequency="4">Johnson</entry>
        <entry frequency="4">Kaifu</entry>
        <entry frequency="4">Kohl</entry>
        <entry frequency="4">Landsbergis</entry>
        <entry frequency="4">Lukanov</entry>
        <entry frequency="4">Marchant</entry>
        <entry frequency="4">Markovic</entry>
        <entry frequency="4">Moi</entry>
        <entry frequency="4">Mugabe</entry>
        <entry frequency="4">Pazner</entry>
        <entry frequency="4">Primakov</entry>
        <entry frequency="4">Ridley</entry>
        <entry frequency="4">Ryzhkov</entry>
        <entry frequency="4">Sharon</entry>
        <entry frequency="4">Singh</entry>
        <entry frequency="4">Taylor</entry>
        <entry frequency="4">Waddington</entry>
        <entry frequency="4">Wijeratne</entry>
        <entry frequency="3">Beron</entry>
        <entry frequency="3">Bullock</entry>
        <entry frequency="3">Carnogursky</entry>
        <entry frequency="3">Cristiani</entry>
        <entry frequency="3">Delors</entry>
        <entry frequency="3">Eyskens</entry>
        <entry frequency="3">Fitzwater</entry>
        <entry frequency="3">Gandhi</entry>
        <entry frequency="3">Gorbachov</entry>
        <entry frequency="3">Hamilton</entry>
        <entry frequency="3">Howell</entry>
        <entry frequency="3">Jovic</entry>
        <entry frequency="3">Khan</entry>
        <entry frequency="3">Lee</entry>
        <entry frequency="3">Lightman</entry>
        <entry frequency="3">Meacher</entry>
        <entry frequency="3">O'Friel</entry>
        <entry frequency="3">Palmer</entry>
        <entry frequency="3">Qian</entry>
        <entry frequency="3">Rahman</entry>
        <entry frequency="3">Savimbi</entry>
        <entry frequency="3">Sawyer</entry>
        <entry frequency="3">Sedley</entry>
        <entry frequency="3">Sharif</entry>
        <entry frequency="3">Tudjman</entry>
        <entry frequency="3">Vlok</entry>
        <entry frequency="3">Williams</entry>
        <entry frequency="3">Wyatt</entry>
    </substitution>
</mwu_substitution>
<lexical_substitution>
    <substitution relation="NN" position="0">
        <entry id="0">
            <score>1.000999000999001</score>
            <substitute>smith</substitute>
            <arguments>insights luxembourg hempstone
                dublin ec gatt hungary lithuania europe</arguments>
        </entry>
        <entry id="1">
            <score>0.33299175147914645</score>
            <substitute>both</substitute>
            <arguments>hungary lithuania europe</arguments>
        </entry>
        <entry id="2">
            <score>0.3335997335997336</score>
            <substitute>brayne</substitute>
            <arguments>luxembourg dublin ec</arguments>
        </entry>
        <entry id="3">
            <score>0.3335997335997336</score>
            <substitute>clifford</substitute>
            <arguments>ec hungary gatt</arguments>
        </entry>
        <entry id="4">
            <score>0.33314719468565623</score>
            <substitute>ec</substitute>
            <arguments>gatt hungary europe</arguments>
        </entry>
    </substitution>
    <substitution relation="NN" position="1">
        <entry id="0">
            <score>1.000999000999001</score>
            <substitute>smith</substitute>
            <arguments>dodie karyn sylvia clifford
                cyril karen ranger joanne adam norman
                desmond will robin eric j rupert anne
                wayne humphrey andy jeffrey shadow
                e hugh gordon ian roger alan mary jan
                martin peter chris steve tony brian
                mike miss colin research professor</arguments>
        </entry>
        <entry id="1">
            <score>0.14606181228044396</score>
            <substitute>reports</substitute>
```

```
            <arguments>jeffrey clifford chris colin
                norman gordon</arguments>
        </entry>
        <entry id="2">
            <score>0.09821257515457146</score>
            <substitute>davis</substitute>
            <arguments>steve colin roger martin</arguments>
        </entry>
        <entry id="3">
            <score>0.09813403923111819</score>
            <substitute>martin</substitute>
            <arguments>mary brian professor ian</arguments>
        </entry>
        <entry id="4">
            <score>0.07326471437969696</score>
            <substitute>cooper</substitute>
            <arguments>roger ian professor</arguments>
        </entry>
        <entry id="5">
            <score>0.07342877795839468</score>
            <substitute>green</substitute>
            <arguments>alan chris professor</arguments>
        </entry>
        <entry id="6">
            <score>0.07304797200267585</score>
            <substitute>jones</substitute>
            <arguments>anne steve ian</arguments>
        </entry>
        <entry id="7">
            <score>0.07331213115185241</score>
            <substitute>michael</substitute>
            <arguments>andy professor peter</arguments>
        </entry>
        <entry id="8">
            <score>0.07323178028082879</score>
            <substitute>powell</substitute>
            <arguments>colin mike chris</arguments>
        </entry>
        <entry id="9">
            <score>0.07344803860483304</score>
            <substitute>richards</substitute>
            <arguments>steve anne brian</arguments>
        </entry>
        <entry id="10">
            <score>0.07317560488292196</score>
            <substitute>stone</substitute>
            <arguments>norman miss peter</arguments>
        </entry>
        <entry id="11">
            <score>0.07313801181397</score>
            <substitute>walker</substitute>
            <arguments>martin chris peter</arguments>
        </entry>
        <entry id="12">
            <score>0.07313966899635738</score>
            <substitute>wilson</substitute>
            <arguments>gordon brian colin</arguments>
        </entry>
    </substitution>
</lexical_substitution>
```

C.5 brown

```
<mwu_substitution>
    <substitution>
        <context frequencyband="14">envelope</context>
        <entry frequency="12">brown</entry>
        <entry frequency="3">sealed</entry>
    </substitution>
    <substitution>
        <context frequencyband="10">eyes</context>
        <entry frequency="3">beautiful</entry>
        <entry frequency="3">blue</entry>
        <entry frequency="3">brown</entry>
    </substitution>
    <substitution>
        <context frequencyband="10">plant</context>
        <entry frequency="3">Scottish</entry>
        <entry frequency="3">brown</entry>
        <entry frequency="3">coca</entry>
        <entry frequency="3">easiest</entry>
        <entry frequency="3">industrial</entry>
        <entry frequency="3">power</entry>
        <entry frequency="3">yam</entry>
    </substitution>
</mwu_substitution>
<lexical_substitution>
    <substitution relation="NN" position="1">
        <entry id="0">
            <score>1.000999000999001</score>
            <substitute>brown</substitute>
            <arguments>jocelyn murphy louise hazel
                gordon basil billy murray ralph nancy
                derek ron ray charlie jerry bobby greg
                simon william tim dr christopher</arguments>
        </entry>
        <entry id="1">
            <score>0.22696135415037905</score>
            <substitute>reports</substitute>
            <arguments>christopher simon murray tim
                gordon</arguments>
        </entry>
        <entry id="2">
            <score>0.18230057162863486</score>
            <substitute>wilson</substitute>
            <arguments>nancy charlie derek gordon</arguments>
        </entry>
        <entry id="3">
            <score>0.1366214812855129</score>
            <substitute>nicholson</substitute>
            <arguments>ralph christopher dr</arguments>
        </entry>
    </substitution>
</lexical_substitution>
```

C.6 computer

```
<mwu_substitution>
    <substitution>
        <context frequencyband="8">asked</context>
        <entry frequency="3">BBC</entry>
        <entry frequency="3">King</entry>
        <entry frequency="3">Saudis</entry>
        <entry frequency="3">ambassador</entry>
        <entry frequency="3">computer</entry>
    </substitution>
    <substitution>
        <context frequencyband="8">company</context>
        <entry frequency="5">US</entry>
        <entry frequency="5">computer</entry>
        <entry frequency="5">drug</entry>
        <entry frequency="5">first</entry>
        <entry frequency="5">mining</entry>
        <entry frequency="5">second</entry>
        <entry frequency="4">Italian</entry>
        <entry frequency="4">Japanese</entry>
        <entry frequency="4">Soviet</entry>
```

```
    <entry frequency="4">Spanish</entry>
    <entry frequency="4">bus</entry>
    <entry frequency="4">merged</entry>
    <entry frequency="4">telecommunications</entry>
    <entry frequency="3">Thai</entry>
    <entry frequency="3">best</entry>
    <entry frequency="3">construction</entry>
    <entry frequency="3">entire</entry>
    <entry frequency="3">holding</entry>
    <entry frequency="3">insolvent</entry>
    <entry frequency="3">insurance</entry>
    <entry frequency="3">issuing</entry>
    <entry frequency="3">multi-national</entry>
    <entry frequency="3">multinational</entry>
    <entry frequency="3">other</entry>
    <entry frequency="3">photographic</entry>
    <entry frequency="3">record</entry>
    <entry frequency="3">television</entry>
    <entry frequency="3">theatre</entry>
</substitution>
<substitution>
    <context frequencyband="10">controlled</context>
    <entry frequency="7">computer</entry>
    <entry frequency="4">precisely</entry>
    <entry frequency="4">remote</entry>
    <entry frequency="3">carefully</entry>
    <entry frequency="3">company</entry>
    <entry frequency="3">state</entry>
    <entry frequency="3">very</entry>
    <entry frequency="3">world</entry>
</substitution>
<substitution>
    <context frequencyband="8">done</context>
    <entry frequency="27">computer</entry>
    <entry frequency="3">Government</entry>
    <entry frequency="3">army</entry>
</substitution>
<substitution>
    <context frequencyband="13">enables</context>
    <entry frequency="4">computer</entry>
    <entry frequency="4">device</entry>
    <entry frequency="4">government</entry>
</substitution>
<substitution>
    <context frequencyband="10">firm</context>
    <entry frequency="20">computer</entry>
    <entry frequency="5">car</entry>
    <entry frequency="3">accountancy</entry>
</substitution>
<substitution>
    <context frequencyband="9">industry</context>
    <entry frequency="35">computer</entry>
    <entry frequency="11">arms</entry>
    <entry frequency="11">music</entry>
    <entry frequency="11">petro-chemical</entry>
    <entry frequency="10">coal</entry>
    <entry frequency="10">electronics</entry>
    <entry frequency="10">leisure</entry>
    <entry frequency="10">meat</entry>
    <entry frequency="9">chemical</entry>
    <entry frequency="9">mining</entry>
    <entry frequency="9">motor</entry>
    <entry frequency="9">pharmaceutical</entry>
    <entry frequency="8">car</entry>
    <entry frequency="8">machine-building</entry>
    <entry frequency="8">nuclear</entry>
    <entry frequency="8">steel</entry>
    <entry frequency="7">airline</entry>
    <entry frequency="7">banking</entry>
    <entry frequency="7">building</entry>
    <entry frequency="7">defence</entry>
    <entry frequency="7">fishing</entry>
    <entry frequency="7">textile</entry>
    <entry frequency="7">timber</entry>
    <entry frequency="7">tobacco</entry>
    <entry frequency="6">securities</entry>
    <entry frequency="6">waste</entry>
    <entry frequency="5">bio</entry>
    <entry frequency="5">service</entry>
    <entry frequency="5">shipping</entry>
    <entry frequency="5">whole</entry>
    <entry frequency="4">aerospace</entry>
    <entry frequency="4">aviation</entry>
    <entry frequency="4">cattle</entry>
    <entry frequency="4">cotton</entry>
    <entry frequency="4">ferry</entry>
    <entry frequency="4">insurance</entry>
    <entry frequency="4">newspaper</entry>
    <entry frequency="4">particular</entry>
    <entry frequency="4">power</entry>
    <entry frequency="4">record</entry>
    <entry frequency="4">sex</entry>
    <entry frequency="4">soup</entry>
    <entry frequency="4">tea</entry>
    <entry frequency="4">tourism</entry>
    <entry frequency="3">US</entry>
    <entry frequency="3">advertising</entry>
    <entry frequency="3">architectual</entry>
    <entry frequency="3">cable</entry>
    <entry frequency="3">carpet</entry>
    <entry frequency="3">dairy</entry>
    <entry frequency="3">farming</entry>
    <entry frequency="3">financial</entry>
    <entry frequency="3">offshore</entry>
    <entry frequency="3">pop</entry>
    <entry frequency="3">seed</entry>
    <entry frequency="3">shipbuilding</entry>
</substitution>
<substitution>
    <context frequencyband="3">is</context>
    <entry frequency="3">Charter</entry>
    <entry frequency="3">Congress</entry>
    <entry frequency="3">FLN</entry>
    <entry frequency="3">Institute</entry>
    <entry frequency="3">Law</entry>
    <entry frequency="3">Pentateuch</entry>
    <entry frequency="3">Senate</entry>
    <entry frequency="3">administration</entry>
    <entry frequency="3">agenda</entry>
    <entry frequency="3">alliance</entry>
    <entry frequency="3">artist</entry>
    <entry frequency="3">attractor</entry>
    <entry frequency="3">average</entry>
    <entry frequency="3">bank</entry>
    <entry frequency="3">cast</entry>
    <entry frequency="3">center</entry>
    <entry frequency="3">change</entry>
    <entry frequency="3">charter</entry>
    <entry frequency="3">church</entry>
    <entry frequency="3">cold</entry>
    <entry frequency="3">community</entry>
    <entry frequency="3">competition</entry>
    <entry frequency="3">computer</entry>
    <entry frequency="3">conflict</entry>
    <entry frequency="3">context</entry>
    <entry frequency="3">corner</entry>
    <entry frequency="3">court</entry>
    <entry frequency="3">customer</entry>
    <entry frequency="3">danger</entry>
    <entry frequency="3">data</entry>
    <entry frequency="3">debate</entry>
    <entry frequency="3">delay</entry>
    <entry frequency="3">development</entry>
    <entry frequency="3">dialogue</entry>
    <entry frequency="3">document</entry>
    <entry frequency="3">dollar</entry>
    <entry frequency="3">east</entry>
    <entry frequency="3">examination</entry>
    <entry frequency="3">exercise</entry>
    <entry frequency="3">explanation</entry>
    <entry frequency="3">fabric</entry>
    <entry frequency="3">feeling</entry>
    <entry frequency="3">fighting</entry>
    <entry frequency="3">fluid</entry>
    <entry frequency="3">fly</entry>
    <entry frequency="3">fund</entry>
    <entry frequency="3">gate</entry>
    <entry frequency="3">grain</entry>
    <entry frequency="3">greater</entry>
    <entry frequency="3">heck</entry>
    <entry frequency="3">hole</entry>
    <entry frequency="3">implication</entry>
    <entry frequency="3">impression</entry>
    <entry frequency="3">interaction</entry>
    <entry frequency="3">island</entry>
    <entry frequency="3">laws</entry>
    <entry frequency="3">legislation</entry>
    <entry frequency="3">light</entry>
    <entry frequency="3">line</entry>
    <entry frequency="3">liner</entry>
    <entry frequency="3">link</entry>
    <entry frequency="3">machine</entry>
    <entry frequency="3">medium</entry>
    <entry frequency="3">mixture</entry>
    <entry frequency="3">mood</entry>
    <entry frequency="3">movie</entry>
```

```
    <entry frequency="3">narrator</entry>
    <entry frequency="3">night</entry>
    <entry frequency="3">object</entry>
    <entry frequency="3">operator</entry>
    <entry frequency="3">opposition</entry>
    <entry frequency="3">orchestra</entry>
    <entry frequency="3">organism</entry>
    <entry frequency="3">part</entry>
    <entry frequency="3">pattern</entry>
    <entry frequency="3">period</entry>
    <entry frequency="3">polka</entry>
    <entry frequency="3">poll</entry>
    <entry frequency="3">postman</entry>
    <entry frequency="3">power</entry>
    <entry frequency="3">practice</entry>
    <entry frequency="3">press</entry>
    <entry frequency="3">probability</entry>
    <entry frequency="3">profession</entry>
    <entry frequency="3">race</entry>
    <entry frequency="3">range</entry>
    <entry frequency="3">reader</entry>
    <entry frequency="3">region</entry>
    <entry frequency="3">relationship</entry>
    <entry frequency="3">request</entry>
    <entry frequency="3">reverse</entry>
    <entry frequency="3">risk</entry>
    <entry frequency="3">sample</entry>
    <entry frequency="3">scene</entry>
    <entry frequency="3">score</entry>
    <entry frequency="3">seminar</entry>
    <entry frequency="3">series</entry>
    <entry frequency="3">site</entry>
    <entry frequency="3">spring</entry>
    <entry frequency="3">state</entry>
    <entry frequency="3">streets</entry>
    <entry frequency="3">tape</entry>
    <entry frequency="3">team</entry>
    <entry frequency="3">top</entry>
    <entry frequency="3">transaction</entry>
    <entry frequency="3">treaty</entry>
    <entry frequency="3">trip</entry>
    <entry frequency="3">tubing</entry>
    <entry frequency="3">tunnel</entry>
    <entry frequency="3">union</entry>
    <entry frequency="3">university</entry>
    <entry frequency="3">verse</entry>
    <entry frequency="3">worker</entry>
</substitution>
<substitution>
    <context frequencyband="9">largest</context>
    <entry frequency="8">computer</entry>
    <entry frequency="3">record</entry>
    <entry frequency="3">stock-broking</entry>
</substitution>
<substitution>
    <context frequencyband="13">manufacturer</context>
    <entry frequency="14">computer</entry>
    <entry frequency="4">car</entry>
</substitution>
<substitution>
    <context frequencyband="10">model</context>
    <entry frequency="47">computer</entry>
    <entry frequency="8">role</entry>
    <entry frequency="7">working</entry>
    <entry frequency="6">compromise</entry>
    <entry frequency="6">simple</entry>
    <entry frequency="5">good</entry>
    <entry frequency="5">scale</entry>
    <entry frequency="4">structural</entry>
    <entry frequency="4">theoretical</entry>
    <entry frequency="3">demonstration</entry>
    <entry frequency="3">mathematical</entry>
    <entry frequency="3">useful</entry>
</substitution>
<substitution>
    <context frequencyband="11">models</context>
    <entry frequency="11">computer</entry>
    <entry frequency="3">three-dimensional</entry>
</substitution>
<substitution>
    <context frequencyband="12">predictions</context>
    <entry frequency="25">computer</entry>
    <entry frequency="6">theoretical</entry>
    <entry frequency="3">best</entry>
    <entry frequency="3">current</entry>
    <entry frequency="3">the</entry>
</substitution>
<substitution>
    <context frequencyband="11">program</context>
    <entry frequency="52">computer</entry>
    <entry frequency="5">premed</entry>
    <entry frequency="4">whole</entry>
</substitution>
<substitution>
    <context frequencyband="8">programme</context>
    <entry frequency="11">computer</entry>
    <entry frequency="9">BBC</entry>
    <entry frequency="9">radical</entry>
    <entry frequency="9">reform</entry>
    <entry frequency="9">television</entry>
    <entry frequency="8">big</entry>
    <entry frequency="8">detailed</entry>
    <entry frequency="8">joint</entry>
    <entry frequency="8">nuclear</entry>
    <entry frequency="7">comprehensive</entry>
    <entry frequency="7">government</entry>
    <entry frequency="7">massive</entry>
    <entry frequency="7">relief</entry>
    <entry frequency="6">clear</entry>
    <entry frequency="5">TV</entry>
    <entry frequency="5">compromise</entry>
    <entry frequency="5">controversial</entry>
    <entry frequency="5">development</entry>
    <entry frequency="4">crash</entry>
    <entry frequency="4">five-hundred-day</entry>
    <entry frequency="4">full</entry>
    <entry frequency="4">huge</entry>
    <entry frequency="4">major</entry>
    <entry frequency="4">model</entry>
    <entry frequency="4">real</entry>
    <entry frequency="4">rolling</entry>
    <entry frequency="4">wide-ranging</entry>
    <entry frequency="3">coherent</entry>
    <entry frequency="3">further</entry>
    <entry frequency="3">military</entry>
    <entry frequency="3">privatisation</entry>
    <entry frequency="3">research</entry>
    <entry frequency="3">single</entry>
    <entry frequency="3">six-point</entry>
</substitution>
<substitution>
    <context frequencyband="10">revolution</context>
    <entry frequency="8">Thatcher</entry>
    <entry frequency="8">computer</entry>
    <entry frequency="7">Chinese</entry>
    <entry frequency="5">Sandinista</entry>
    <entry frequency="5">anti-Communist</entry>
    <entry frequency="5">green</entry>
    <entry frequency="5">political</entry>
    <entry frequency="4">Ethiopian</entry>
    <entry frequency="4">Socialist-based</entry>
    <entry frequency="4">Soviet</entry>
    <entry frequency="4">popular</entry>
    <entry frequency="4">scientific</entry>
    <entry frequency="3">Cuban</entry>
    <entry frequency="3">Czechoslovak</entry>
    <entry frequency="3">French</entry>
    <entry frequency="3">November</entry>
    <entry frequency="3">Polish</entry>
    <entry frequency="3">anti-communist</entry>
    <entry frequency="3">bloody</entry>
    <entry frequency="3">communist</entry>
    <entry frequency="3">counter</entry>
    <entry frequency="3">cultural</entry>
    <entry frequency="3">market</entry>
</substitution>
<substitution>
    <context frequencyband="9">room</context>
    <entry frequency="3">common</entry>
    <entry frequency="3">computer</entry>
    <entry frequency="3">cutting</entry>
    <entry frequency="3">empty</entry>
    <entry frequency="3">incident</entry>
    <entry frequency="3">shower</entry>
    <entry frequency="3">small</entry>
    <entry frequency="3">strong</entry>
    <entry frequency="3">throne</entry>
    <entry frequency="3">utility</entry>
</substitution>
<substitution>
    <context frequencyband="4">said</context>
    <entry frequency="10">King</entry>
    <entry frequency="10">computer</entry>
    <entry frequency="10">official</entry>
    <entry frequency="9">president</entry>
    <entry frequency="7">authorities</entry>
```

```
    <entry frequency="7">judge</entry>
    <entry frequency="7">woman</entry>
    <entry frequency="6">OD</entry>
    <entry frequency="6">report</entry>
    <entry frequency="5">IRA</entry>
    <entry frequency="5">army</entry>
    <entry frequency="5">girl</entry>
    <entry frequency="5">meeting</entry>
    <entry frequency="5">spokeswoman</entry>
    <entry frequency="5">statement</entry>
    <entry frequency="4">driver</entry>
    <entry frequency="4">ex-soldier</entry>
    <entry frequency="4">man</entry>
    <entry frequency="4">organisers</entry>
    <entry frequency="4">sheriff</entry>
    <entry frequency="4">shopkeeper</entry>
    <entry frequency="3">Archbishop</entry>
    <entry frequency="3">Minister</entry>
    <entry frequency="3">ambassador</entry>
    <entry frequency="3">commission</entry>
    <entry frequency="3">diplomat</entry>
    <entry frequency="3">experts</entry>
    <entry frequency="3">general</entry>
    <entry frequency="3">group</entry>
    <entry frequency="3">plant</entry>
    <entry frequency="3">poet</entry>
    <entry frequency="3">rebels</entry>
    <entry frequency="3">source</entry>
</substitution>
<substitution>
    <context frequencyband="12">screen</context>
    <entry frequency="9">computer</entry>
    <entry frequency="3">large</entry>
    <entry frequency="3">television</entry>
</substitution>
<substitution>
    <context frequencyband="14">screens</context>
    <entry frequency="51">computer</entry>
    <entry frequency="3">television</entry>
    <entry frequency="3">the</entry>
</substitution>
<substitution>
    <context frequencyband="7">system</context>
    <entry frequency="51">computer</entry>
    <entry frequency="22">federal</entry>
    <entry frequency="22">financial</entry>
    <entry frequency="21">economic</entry>
    <entry frequency="20">communist</entry>
    <entry frequency="19">distribution</entry>
    <entry frequency="18">regimental</entry>
    <entry frequency="16">market</entry>
    <entry frequency="16">panchayat</entry>
    <entry frequency="15">controversial</entry>
    <entry frequency="15">previous</entry>
    <entry frequency="15">sub</entry>
    <entry frequency="14">presidential</entry>
    <entry frequency="13">democratic</entry>
    <entry frequency="12">telephone</entry>
    <entry frequency="12">two-party</entry>
    <entry frequency="11">complex</entry>
    <entry frequency="11">parliamentary</entry>
    <entry frequency="11">planning</entry>
    <entry frequency="11">single-party</entry>
    <entry frequency="11">state</entry>
    <entry frequency="10">British</entry>
    <entry frequency="10">climate</entry>
    <entry frequency="9">educational</entry>
    <entry frequency="9">legal</entry>
    <entry frequency="9">loudspeaker</entry>
    <entry frequency="9">railway</entry>
    <entry frequency="9">socialist</entry>
    <entry frequency="8">Lorenz</entry>
    <entry frequency="8">Panchayat</entry>
    <entry frequency="8">caste</entry>
    <entry frequency="8">quota</entry>
    <entry frequency="8">simple</entry>
    <entry frequency="7">best</entry>
    <entry frequency="7">capitalist</entry>
    <entry frequency="7">class</entry>
    <entry frequency="7">just</entry>
    <entry frequency="7">monetary</entry>
    <entry frequency="7">planetary</entry>
    <entry frequency="7">social</entry>
    <entry frequency="7">transport</entry>
    <entry frequency="6">banking</entry>
    <entry frequency="6">closed</entry>
    <entry frequency="6">comprehensive</entry>
    <entry frequency="6">defence</entry>
    <entry frequency="6">international</entry>
    <entry frequency="6">other</entry>
    <entry frequency="6">penal</entry>
    <entry frequency="6">same</entry>
    <entry frequency="6">sewerage</entry>
    <entry frequency="6">similar</entry>
    <entry frequency="6">unique</entry>
    <entry frequency="6">whole</entry>
    <entry frequency="5">French</entry>
    <entry frequency="5">Socialist</entry>
    <entry frequency="5">adversary</entry>
    <entry frequency="5">binary</entry>
    <entry frequency="5">cAMP</entry>
    <entry frequency="5">complete</entry>
    <entry frequency="5">former</entry>
    <entry frequency="5">new</entry>
    <entry frequency="5">non-party</entry>
    <entry frequency="5">screening</entry>
    <entry frequency="5">sprinkler</entry>
    <entry frequency="5">two-tier</entry>
    <entry frequency="5">university</entry>
    <entry frequency="5">ventilation</entry>
    <entry frequency="5">voluntary</entry>
    <entry frequency="4">American</entry>
    <entry frequency="4">European</entry>
    <entry frequency="4">UN</entry>
    <entry frequency="4">blood</entry>
    <entry frequency="4">confederal</entry>
    <entry frequency="4">control</entry>
    <entry frequency="4">dissipative</entry>
    <entry frequency="4">education</entry>
    <entry frequency="4">entire</entry>
    <entry frequency="4">good</entry>
    <entry frequency="4">homeland</entry>
    <entry frequency="4">honours</entry>
    <entry frequency="4">hydraulic</entry>
    <entry frequency="4">information</entry>
    <entry frequency="4">justice</entry>
    <entry frequency="4">licensing</entry>
    <entry frequency="4">lymphatic</entry>
    <entry frequency="4">party-list</entry>
    <entry frequency="4">payment</entry>
    <entry frequency="4">practical</entry>
    <entry frequency="4">protective</entry>
    <entry frequency="4">totalitarian</entry>
    <entry frequency="4">workable</entry>
    <entry frequency="3">Australian</entry>
    <entry frequency="3">Community</entry>
    <entry frequency="3">Himalayan</entry>
    <entry frequency="3">Parliamentary</entry>
    <entry frequency="3">Soviet</entry>
    <entry frequency="3">automatic</entry>
    <entry frequency="3">basic</entry>
    <entry frequency="3">better</entry>
    <entry frequency="3">cave</entry>
    <entry frequency="3">competitive</entry>
    <entry frequency="3">current</entry>
    <entry frequency="3">dynamical</entry>
    <entry frequency="3">earth-atmosphere</entry>
    <entry frequency="3">electoral</entry>
    <entry frequency="3">established</entry>
    <entry frequency="3">evil</entry>
    <entry frequency="3">fairer</entry>
    <entry frequency="3">feudal</entry>
    <entry frequency="3">first-past-the-post</entry>
    <entry frequency="3">imperial</entry>
    <entry frequency="3">judicial</entry>
    <entry frequency="3">jury</entry>
    <entry frequency="3">language</entry>
    <entry frequency="3">marketing</entry>
    <entry frequency="3">model</entry>
    <entry frequency="3">motor</entry>
    <entry frequency="3">nervous</entry>
    <entry frequency="3">party</entry>
    <entry frequency="3">partyless</entry>
    <entry frequency="3">perceptual</entry>
    <entry frequency="3">phone</entry>
    <entry frequency="3">political</entry>
    <entry frequency="3">prison</entry>
    <entry frequency="3">proposed</entry>
    <entry frequency="3">radar</entry>
    <entry frequency="3">rail</entry>
    <entry frequency="3">rating</entry>
    <entry frequency="3">rational</entry>
    <entry frequency="3">security</entry>
    <entry frequency="3">sewage</entry>
    <entry frequency="3">shift</entry>
    <entry frequency="3">sign</entry>
```

```
    <entry frequency="3">taxation</entry>
    <entry frequency="3">total</entry>
    <entry frequency="3">voting</entry>
    <entry frequency="3">voucher</entry>
</substitution>
<substitution>
    <context frequencyband="10">systems</context>
    <entry frequency="13">computer</entry>
    <entry frequency="3">mountain</entry>
    <entry frequency="3">one-party</entry>
    <entry frequency="3">the</entry>
    <entry frequency="3">these</entry>
    <entry frequency="3">two</entry>
</substitution>
<substitution>
    <context frequencyband="1">to</context>
    <entry frequency="3">Arabs</entry>
    <entry frequency="3">Atlantic</entry>
    <entry frequency="3">Baltic</entry>
    <entry frequency="3">British</entry>
    <entry frequency="3">French</entry>
    <entry frequency="3">Tories</entry>
    <entry frequency="3">UK</entry>
    <entry frequency="3">advantages</entry>
    <entry frequency="3">appeal</entry>
    <entry frequency="3">approaches</entry>
    <entry frequency="3">army</entry>
    <entry frequency="3">assurance</entry>
    <entry frequency="3">attempts</entry>
    <entry frequency="3">award</entry>
    <entry frequency="3">barriers</entry>
    <entry frequency="3">benefit</entry>
    <entry frequency="3">bid</entry>
    <entry frequency="3">boom</entry>
    <entry frequency="3">camera</entry>
    <entry frequency="3">cash</entry>
    <entry frequency="3">centre</entry>
    <entry frequency="3">challenge</entry>
    <entry frequency="3">check</entry>
    <entry frequency="3">city</entry>
    <entry frequency="3">commission</entry>
    <entry frequency="3">commonalty</entry>
    <entry frequency="3">computer</entry>
    <entry frequency="3">confidence</entry>
    <entry frequency="3">contract</entry>
    <entry frequency="3">contras</entry>
    <entry frequency="3">cost</entry>
    <entry frequency="3">edges</entry>
    <entry frequency="3">facilities</entry>
    <entry frequency="3">fight</entry>
    <entry frequency="3">fighting</entry>
    <entry frequency="3">forest</entry>
    <entry frequency="3">heading</entry>
    <entry frequency="3">heir</entry>
    <entry frequency="3">interpreters</entry>
    <entry frequency="3">issue</entry>
    <entry frequency="3">kitchen</entry>
    <entry frequency="3">known</entry>
    <entry frequency="3">liner</entry>
    <entry frequency="3">lyrics</entry>
    <entry frequency="3">man</entry>
    <entry frequency="3">material</entry>
    <entry frequency="3">medium</entry>
    <entry frequency="3">military</entry>
    <entry frequency="3">milk</entry>
    <entry frequency="3">miners</entry>
    <entry frequency="3">months</entry>
    <entry frequency="3">music</entry>
    <entry frequency="3">necessity</entry>
    <entry frequency="3">obstacles</entry>
    <entry frequency="3">option</entry>
    <entry frequency="3">parties</entry>
    <entry frequency="3">past</entry>
    <entry frequency="3">path</entry>
    <entry frequency="3">picture</entry>
    <entry frequency="3">plane</entry>
    <entry frequency="3">plot</entry>
    <entry frequency="3">podium</entry>
    <entry frequency="3">preface</entry>
    <entry frequency="3">presidency</entry>
    <entry frequency="3">pressure</entry>
    <entry frequency="3">price</entry>
    <entry frequency="3">prison</entry>
    <entry frequency="3">problem</entry>
    <entry frequency="3">proposal</entry>
    <entry frequency="3">proposals</entry>
    <entry frequency="3">results</entry>
    <entry frequency="3">rights</entry>
    <entry frequency="3">river</entry>
    <entry frequency="3">rules</entry>
    <entry frequency="3">sand</entry>
    <entry frequency="3">situation</entry>
    <entry frequency="3">soldiers</entry>
    <entry frequency="3">solution</entry>
    <entry frequency="3">stage</entry>
    <entry frequency="3">struggle</entry>
    <entry frequency="3">sum</entry>
    <entry frequency="3">sun</entry>
    <entry frequency="3">surface</entry>
    <entry frequency="3">tendency</entry>
    <entry frequency="3">text</entry>
    <entry frequency="3">wisdom</entry>
    <entry frequency="3">wish</entry>
    <entry frequency="3">women</entry>
</substitution>
<substitution>
    <context frequencyband="6">world</context>
    <entry frequency="3">adult</entry>
    <entry frequency="3">civilized</entry>
    <entry frequency="3">competitive</entry>
    <entry frequency="3">complex</entry>
    <entry frequency="3">computer</entry>
    <entry frequency="3">contemporary</entry>
    <entry frequency="3">created</entry>
    <entry frequency="3">defending</entry>
    <entry frequency="3">forthcoming</entry>
    <entry frequency="3">greenhouse</entry>
    <entry frequency="3">inanimate</entry>
    <entry frequency="3">medical</entry>
    <entry frequency="3">only</entry>
    <entry frequency="3">postwar</entry>
    <entry frequency="3">round-the</entry>
</substitution>
</mwu_substitution>
<lexical_substitution>
    <substitution relation="NN" position="0">
        <entry id="0">
            <score>1.0009990009990001</score>
            <substitute>computer</substitute>
            <arguments>hackers projections simulations
                simulation database predictions terminals
                software screens models manual screen
                crime memories manufacturer array chips
                corp technology data program maker
                images memory model fraud programs
                files crimes applications science records
                equipment analysis command studies
                scientists</arguments>
        </entry>
        <entry id="1">
            <score>0.13500522018421368</score>
            <substitute>your</substitute>
            <arguments>simulations memory predictions
                data model</arguments>
        </entry>
        <entry id="2">
            <score>0.10805120436699384</score>
            <substitute>space</substitute>
            <arguments>images scientists technology
                science</arguments>
        </entry>
        <entry id="3">
            <score>0.08104972926704676</score>
            <substitute>animal</substitute>
            <arguments>models studies model</arguments>
        </entry>
        <entry id="4">
            <score>0.08110222563663859</score>
            <substitute>car</substitute>
            <arguments>maker manufacturer crime</arguments>
        </entry>
        <entry id="5">
            <score>0.0810099047194189</score>
            <substitute>french</substitute>
            <arguments>manufacturer model scientists</arguments>
        </entry>
        <entry id="6">
            <score>0.08105163866660632</score>
            <substitute>japanese</substitute>
            <arguments>manufacturer scientists technology</arguments>
        </entry>
        <entry id="7">
            <score>0.08102651464788149</score>
            <substitute>military</substitute>
            <arguments>command equipment technology</arguments>
        </entry>
```

```
        <entry id="8">
            <score>0.0809874469831843</score>
            <substitute>my</substitute>
            <arguments>memory memories studies</arguments>
        </entry>
        <entry id="9">
            <score>0.08100439153070732</score>
            <substitute>our</substitute>
            <arguments>science simulations data</arguments>
        </entry>
        <entry id="10">
            <score>0.08144496565549197</score>
            <substitute>radar</substitute>
            <arguments>screen screens equipment</arguments>
        </entry>
        <entry id="11">
            <score>0.08116023379181274</score>
            <substitute>satellite</substitute>
            <arguments>images data studies</arguments>
        </entry>
    </substitution>
    <substitution relation="NN" position="1">
        <entry id="0">
            <score>1.000999000999001</score>
            <substitute>computer</substitute>
            <arguments>reality oed japanese your home
                britain police my</arguments>
        </entry>
        <entry id="1">
            <score>0.37553557553557554</score>
            <substitute>stores</substitute>
            <arguments>home britain your</arguments>
        </entry>
    </substitution>
</lexical_substitution>
```

C.7 dry

```
<mwu_substitution>
    <substitution>
        <context frequencyband="8">allowed</context>
        <entry frequency="3">come</entry>
        <entry frequency="3">dry</entry>
    </substitution>
    <substitution>
        <context frequencyband="8">area</context>
        <entry frequency="14">dry</entry>
        <entry frequency="8">large</entry>
        <entry frequency="6">complex</entry>
        <entry frequency="6">different</entry>
        <entry frequency="3">fertile</entry>
        <entry frequency="3">small</entry>
    </substitution>
    <substitution>
        <context frequencyband="8">areas</context>
        <entry frequency="10">dry</entry>
        <entry frequency="7">northern</entry>
        <entry frequency="7">tropical</entry>
        <entry frequency="6">Kurdish</entry>
        <entry frequency="6">four</entry>
        <entry frequency="6">sensitive</entry>
        <entry frequency="6">troubled</entry>
        <entry frequency="5">Arab</entry>
        <entry frequency="5">contaminated</entry>
        <entry frequency="5">marginal</entry>
        <entry frequency="5">neighbouring</entry>
        <entry frequency="4">Muslim</entry>
        <entry frequency="4">both</entry>
        <entry frequency="4">built-up</entry>
        <entry frequency="4">coastal</entry>
        <entry frequency="4">designated</entry>
        <entry frequency="4">important</entry>
        <entry frequency="4">jungle</entry>
        <entry frequency="4">specific</entry>
        <entry frequency="4">strategic</entry>
        <entry frequency="4">vital</entry>
        <entry frequency="3">arid</entry>
        <entry frequency="3">country</entry>
        <entry frequency="3">crowded</entry>
        <entry frequency="3">deep</entry>
        <entry frequency="3">industrial</entry>
        <entry frequency="3">mixed</entry>
        <entry frequency="3">nearby</entry>
        <entry frequency="3">patient</entry>
        <entry frequency="3">separate</entry>
        <entry frequency="3">slum</entry>
        <entry frequency="3">small</entry>
        <entry frequency="3">to</entry>
        <entry frequency="3">uninhabited</entry>
        <entry frequency="3">wet</entry>
    </substitution>
    <substitution>
        <context frequencyband="4">as</context>
        <entry frequency="4">dry</entry>
        <entry frequency="4">far</entry>
        <entry frequency="4">white</entry>
        <entry frequency="3">late</entry>
        <entry frequency="3">much</entry>
        <entry frequency="3">tall</entry>
    </substitution>
    <substitution>
        <context frequencyband="9">land</context>
        <entry frequency="3">acquire</entry>
        <entry frequency="3">building</entry>
        <entry frequency="3">clear</entry>
        <entry frequency="3">dry</entry>
        <entry frequency="3">inherit</entry>
        <entry frequency="3">leave</entry>
        <entry frequency="3">own</entry>
        <entry frequency="3">redistribute</entry>
        <entry frequency="3">use</entry>
    </substitution>
    <substitution>
        <context frequencyband="5">out</context>
        <entry frequency="10">dry</entry>
        <entry frequency="6">bail</entry>
        <entry frequency="6">bow</entry>
        <entry frequency="6">check</entry>
        <entry frequency="6">climb</entry>
        <entry frequency="6">dig</entry>
        <entry frequency="6">drive</entry>
        <entry frequency="6">miss</entry>
        <entry frequency="6">throw</entry>
        <entry frequency="6">weed</entry>
        <entry frequency="5">buy</entry>
        <entry frequency="5">drag</entry>
        <entry frequency="5">drop</entry>
        <entry frequency="5">hand</entry>
        <entry frequency="5">jump</entry>
        <entry frequency="5">knock</entry>
        <entry frequency="5">lash</entry>
        <entry frequency="5">leave</entry>
        <entry frequency="5">look</entry>
        <entry frequency="5">thrash</entry>
        <entry frequency="4">back</entry>
        <entry frequency="4">come</entry>
        <entry frequency="4">draw</entry>
        <entry frequency="4">flesh</entry>
        <entry frequency="4">flush</entry>
        <entry frequency="4">get</entry>
        <entry frequency="4">let</entry>
        <entry frequency="4">ride</entry>
        <entry frequency="4">search</entry>
        <entry frequency="4">sell</entry>
        <entry frequency="4">send</entry>
        <entry frequency="4">stretch</entry>
        <entry frequency="4">tease</entry>
        <entry frequency="4">working</entry>
        <entry frequency="3">bear</entry>
        <entry frequency="3">block</entry>
        <entry frequency="3">drown</entry>
        <entry frequency="3">eat</entry>
        <entry frequency="3">eke</entry>
        <entry frequency="3">fall</entry>
        <entry frequency="3">flatten</entry>
        <entry frequency="3">give</entry>
```

```
        <entry frequency="3">hang</entry>
        <entry frequency="3">lay</entry>
        <entry frequency="3">leak</entry>
        <entry frequency="3">lose</entry>
        <entry frequency="3">pump</entry>
        <entry frequency="3">single</entry>
        <entry frequency="3">sit</entry>
        <entry frequency="3">smooth</entry>
        <entry frequency="3">stand</entry>
        <entry frequency="3">test</entry>
        <entry frequency="3">travel</entry>
    </substitution>
    <substitution>
        <context frequencyband="10">season</context>
        <entry frequency="50">dry</entry>
        <entry frequency="18">holiday</entry>
        <entry frequency="14">football</entry>
        <entry frequency="12">English</entry>
        <entry frequency="12">off</entry>
        <entry frequency="12">tourist</entry>
        <entry frequency="11">close</entry>
        <entry frequency="8">winter</entry>
        <entry frequency="7">festive</entry>
        <entry frequency="6">growing</entry>
        <entry frequency="6">rainy</entry>
        <entry frequency="5">Christmas</entry>
        <entry frequency="5">coming</entry>
        <entry frequency="4">flood</entry>
        <entry frequency="4">regular</entry>
        <entry frequency="4">summer</entry>
        <entry frequency="4">wet</entry>
        <entry frequency="3">current</entry>
        <entry frequency="3">high</entry>
        <entry frequency="3">hunting</entry>
        <entry frequency="3">league</entry>
        <entry frequency="3">lean</entry>
        <entry frequency="3">monsoon</entry>
        <entry frequency="3">sailing</entry>
    </substitution>
    <substitution>
        <context frequencyband="10">summer</context>
        <entry frequency="3">dry</entry>
        <entry frequency="3">fine</entry>
        <entry frequency="3">new</entry>
    </substitution>
    <substitution>
        <context frequencyband="5">up</context>
        <entry frequency="3">bring</entry>
        <entry frequency="3">draw</entry>
        <entry frequency="3">dry</entry>
        <entry frequency="3">grow</entry>
        <entry frequency="3">stand</entry>
        <entry frequency="3">stay</entry>
        <entry frequency="3">travel</entry>
    </substitution>
    <substitution>
        <context frequencyband="3">was</context>
        <entry frequency="4">arrested</entry>
        <entry frequency="4">bored</entry>
        <entry frequency="4">closed</entry>
        <entry frequency="4">discontinued</entry>
        <entry frequency="4">dry</entry>
        <entry frequency="4">found</entry>
        <entry frequency="4">happening</entry>
        <entry frequency="4">hot</entry>
        <entry frequency="4">ill</entry>
        <entry frequency="4">shot</entry>
        <entry frequency="4">simple</entry>
        <entry frequency="4">stabbed</entry>
        <entry frequency="4">weak</entry>
        <entry frequency="4">well</entry>
        <entry frequency="3">attacked</entry>
        <entry frequency="3">black</entry>
        <entry frequency="3">bright</entry>
        <entry frequency="3">calm</entry>
        <entry frequency="3">captured</entry>
        <entry frequency="3">clear</entry>
        <entry frequency="3">crying</entry>
        <entry frequency="3">cut</entry>
        <entry frequency="3">eight</entry>
        <entry frequency="3">formed</entry>
        <entry frequency="3">gone</entry>
        <entry frequency="3">informed</entry>
        <entry frequency="3">launched</entry>
        <entry frequency="3">long</entry>
        <entry frequency="3">made</entry>
        <entry frequency="3">necessary</entry>
        <entry frequency="3">real</entry>
        <entry frequency="3">rich</entry>
        <entry frequency="3">said</entry>
        <entry frequency="3">saying</entry>
        <entry frequency="3">short</entry>
        <entry frequency="3">swift</entry>
        <entry frequency="3">thick</entry>
        <entry frequency="3">tried</entry>
        <entry frequency="3">working</entry>
        <entry frequency="3">wounded</entry>
    </substitution>
    <substitution>
        <context frequencyband="11">weight</context>
        <entry frequency="6">dry</entry>
        <entry frequency="3">full</entry>
        <entry frequency="3">political</entry>
    </substitution>
</mwu_substitution>
<lexical_substitution>
</lexical_substitution>
```

C.8 plant

```
<mwu_substitution>
    <substitution>
        <context frequencyband="7">another</context>
        <entry frequency="3">authority</entry>
        <entry frequency="3">day</entry>
        <entry frequency="3">generation</entry>
        <entry frequency="3">person</entry>
        <entry frequency="3">plant</entry>
        <entry frequency="3">point</entry>
    </substitution>
    <substitution>
        <context frequencyband="4">at</context>
        <entry frequency="3">holes</entry>
        <entry frequency="3">plant</entry>
        <entry frequency="3">summit</entry>
        <entry frequency="3">weekend</entry>
    </substitution>
    <substitution>
        <context frequencyband="9">bomb</context>
        <entry frequency="4">plant</entry>
        <entry frequency="3">be</entry>
        <entry frequency="3">defuse</entry>
    </substitution>
    <substitution>
        <context frequencyband="9">chemical</context>
        <entry frequency="28">plant</entry>
        <entry frequency="6">factory</entry>
        <entry frequency="4">reactions</entry>
        <entry frequency="3">assistant</entry>
        <entry frequency="3">complex</entry>
    </substitution>
    <substitution>
        <context frequencyband="8">life</context>
        <entry frequency="14">family</entry>
        <entry frequency="14">plant</entry>
        <entry frequency="14">your</entry>
        <entry frequency="12">modern</entry>
        <entry frequency="12">what</entry>
        <entry frequency="10">marine</entry>
        <entry frequency="10">national</entry>
        <entry frequency="10">normal</entry>
        <entry frequency="9">British</entry>
        <entry frequency="9">daily</entry>
        <entry frequency="7">healthy</entry>
        <entry frequency="6">our</entry>
        <entry frequency="6">public</entry>
```

```
    <entry frequency="6">rural</entry>
    <entry frequency="6">social</entry>
    <entry frequency="6">this</entry>
    <entry frequency="5">new</entry>
    <entry frequency="5">ordinary</entry>
    <entry frequency="5">private</entry>
    <entry frequency="4">economic</entry>
    <entry frequency="4">parliamentary</entry>
    <entry frequency="3">all</entry>
    <entry frequency="3">prolonging</entry>
    <entry frequency="3">supporting</entry>
    <entry frequency="3">thy</entry>
</substitution>
<substitution>
    <context frequencyband="8">nuclear</context>
    <entry frequency="60">plant</entry>
    <entry frequency="13">plants</entry>
    <entry frequency="10">station</entry>
    <entry frequency="9">stations</entry>
    <entry frequency="8">disaster</entry>
    <entry frequency="7">accidents</entry>
    <entry frequency="7">equipment</entry>
    <entry frequency="6">deterrent</entry>
    <entry frequency="6">facilities</entry>
    <entry frequency="5">power</entry>
    <entry frequency="4">presence</entry>
    <entry frequency="4">warheads</entry>
    <entry frequency="3">accident</entry>
    <entry frequency="3">arsenal</entry>
    <entry frequency="3">artillery</entry>
    <entry frequency="3">device</entry>
    <entry frequency="3">experts</entry>
    <entry frequency="3">fusion</entry>
    <entry frequency="3">lobby</entry>
    <entry frequency="3">materials</entry>
    <entry frequency="3">movement</entry>
    <entry frequency="3">non-proliferation</entry>
    <entry frequency="3">proliferation</entry>
    <entry frequency="3">strategy</entry>
    <entry frequency="3">test</entry>
    <entry frequency="3">tests</entry>
</substitution>
<substitution>
    <context frequencyband="7">power</context>
    <entry frequency="25">plant</entry>
    <entry frequency="13">struggle</entry>
    <entry frequency="12">structure</entry>
    <entry frequency="11">summit</entry>
    <entry frequency="9">plants</entry>
    <entry frequency="7">but</entry>
    <entry frequency="7">rights</entry>
    <entry frequency="6">vacuum</entry>
    <entry frequency="5">and</entry>
    <entry frequency="5">project</entry>
    <entry frequency="5">station</entry>
    <entry frequency="4">is</entry>
    <entry frequency="4">talks</entry>
    <entry frequency="4">was</entry>
    <entry frequency="3">failure</entry>
    <entry frequency="3">relationships</entry>
    <entry frequency="3">were</entry>
</substitution>
<substitution>
    <context frequencyband="3">was</context>
    <entry frequency="5">Church</entry>
    <entry frequency="5">Russians</entry>
    <entry frequency="5">atmosphere</entry>
    <entry frequency="5">band</entry>
    <entry frequency="5">business</entry>
    <entry frequency="5">charge</entry>
    <entry frequency="5">court</entry>
    <entry frequency="5">deceased</entry>
    <entry frequency="5">fighting</entry>
    <entry frequency="5">film</entry>
    <entry frequency="5">fire</entry>
    <entry frequency="5">firm</entry>
    <entry frequency="5">hospital</entry>
    <entry frequency="5">incident</entry>
    <entry frequency="5">letter</entry>
    <entry frequency="5">market</entry>
    <entry frequency="5">marriage</entry>
    <entry frequency="5">massacre</entry>
    <entry frequency="5">name</entry>
    <entry frequency="5">operation</entry>
    <entry frequency="5">organisation</entry>
    <entry frequency="5">place</entry>
    <entry frequency="5">plant</entry>
    <entry frequency="5">process</entry>
    <entry frequency="5">ship</entry>
    <entry frequency="5">strike</entry>
    <entry frequency="5">study</entry>
    <entry frequency="5">subject</entry>
    <entry frequency="5">truth</entry>
    <entry frequency="5">unit</entry>
    <entry frequency="5">word</entry>
    <entry frequency="4">IMF</entry>
    <entry frequency="4">PLO</entry>
    <entry frequency="4">Underworld</entry>
    <entry frequency="4">answer</entry>
    <entry frequency="4">arrangement</entry>
    <entry frequency="4">bill</entry>
    <entry frequency="4">body</entry>
    <entry frequency="4">bomb</entry>
    <entry frequency="4">campaign</entry>
    <entry frequency="4">church</entry>
    <entry frequency="4">conversation</entry>
    <entry frequency="4">discussion</entry>
    <entry frequency="4">election</entry>
    <entry frequency="4">escape</entry>
    <entry frequency="4">event</entry>
    <entry frequency="4">fund</entry>
    <entry frequency="4">garden</entry>
    <entry frequency="4">hell</entry>
    <entry frequency="4">hope</entry>
    <entry frequency="4">intention</entry>
    <entry frequency="4">land</entry>
    <entry frequency="4">law</entry>
    <entry frequency="4">loan</entry>
    <entry frequency="4">match</entry>
    <entry frequency="4">measure</entry>
    <entry frequency="4">movement</entry>
    <entry frequency="4">news</entry>
    <entry frequency="4">offence</entry>
    <entry frequency="4">original</entry>
    <entry frequency="4">outcome</entry>
    <entry frequency="4">paper</entry>
    <entry frequency="4">play</entry>
    <entry frequency="4">population</entry>
    <entry frequency="4">pound</entry>
    <entry frequency="4">prince</entry>
    <entry frequency="4">prison</entry>
    <entry frequency="4">procedure</entry>
    <entry frequency="4">programme</entry>
    <entry frequency="4">public</entry>
    <entry frequency="4">reason</entry>
    <entry frequency="4">region</entry>
    <entry frequency="4">site</entry>
    <entry frequency="4">sky</entry>
    <entry frequency="4">solution</entry>
    <entry frequency="4">sport</entry>
    <entry frequency="4">summit</entry>
    <entry frequency="4">survey</entry>
    <entry frequency="4">table</entry>
    <entry frequency="4">talks</entry>
    <entry frequency="4">text</entry>
    <entry frequency="4">tunnel</entry>
    <entry frequency="4">victim</entry>
    <entry frequency="3">BBC</entry>
    <entry frequency="3">CDU</entry>
    <entry frequency="3">Colonel</entry>
    <entry frequency="3">Commons</entry>
    <entry frequency="3">Community</entry>
    <entry frequency="3">Duke</entry>
    <entry frequency="3">EC</entry>
    <entry frequency="3">East</entry>
    <entry frequency="3">NHS</entry>
    <entry frequency="3">Rector</entry>
    <entry frequency="3">Treaty</entry>
    <entry frequency="3">Universe</entry>
    <entry frequency="3">aid</entry>
    <entry frequency="3">aim</entry>
    <entry frequency="3">animal</entry>
    <entry frequency="3">appeal</entry>
    <entry frequency="3">baby</entry>
    <entry frequency="3">ball</entry>
    <entry frequency="3">bed</entry>
    <entry frequency="3">century</entry>
    <entry frequency="3">ceremony</entry>
    <entry frequency="3">child</entry>
    <entry frequency="3">circumstances</entry>
    <entry frequency="3">claim</entry>
    <entry frequency="3">coast</entry>
    <entry frequency="3">codex</entry>
    <entry frequency="3">congress</entry>
    <entry frequency="3">continent</entry>
    <entry frequency="3">crew</entry>
```

```
            <entry frequency="3">crime</entry>
            <entry frequency="3">crisis</entry>
            <entry frequency="3">criticism</entry>
            <entry frequency="3">curfew</entry>
            <entry frequency="3">damage</entry>
            <entry frequency="3">debate</entry>
            <entry frequency="3">deficit</entry>
            <entry frequency="3">development</entry>
            <entry frequency="3">difference</entry>
            <entry frequency="3">disease</entry>
            <entry frequency="3">dog</entry>
            <entry frequency="3">driver</entry>
            <entry frequency="3">emphasis</entry>
            <entry frequency="3">engine</entry>
            <entry frequency="3">episode</entry>
            <entry frequency="3">equipment</entry>
            <entry frequency="3">evidence</entry>
            <entry frequency="3">explosion</entry>
            <entry frequency="3">farm</entry>
            <entry frequency="3">festival</entry>
            <entry frequency="3">figure</entry>
            <entry frequency="3">front</entry>
            <entry frequency="3">frontier</entry>
            <entry frequency="3">fuss</entry>
            <entry frequency="3">general</entry>
            <entry frequency="3">gloom</entry>
            <entry frequency="3">group</entry>
            <entry frequency="3">hall</entry>
            <entry frequency="3">investigation</entry>
            <entry frequency="3">kitchen</entry>
            <entry frequency="3">language</entry>
            <entry frequency="3">legislation</entry>
            <entry frequency="3">light</entry>
            <entry frequency="3">military</entry>
            <entry frequency="3">mission</entry>
            <entry frequency="3">mood</entry>
            <entry frequency="3">mosque</entry>
            <entry frequency="3">murder</entry>
            <entry frequency="3">music</entry>
            <entry frequency="3">night</entry>
            <entry frequency="3">number</entry>
            <entry frequency="3">office</entry>
            <entry frequency="3">paint</entry>
            <entry frequency="3">past</entry>
            <entry frequency="3">performance</entry>
            <entry frequency="3">person</entry>
            <entry frequency="3">phone</entry>
            <entry frequency="3">picture</entry>
            <entry frequency="3">presidency</entry>
            <entry frequency="3">president</entry>
            <entry frequency="3">property</entry>
            <entry frequency="3">radio</entry>
            <entry frequency="3">range</entry>
            <entry frequency="3">request</entry>
            <entry frequency="3">review</entry>
            <entry frequency="3">right</entry>
            <entry frequency="3">river</entry>
            <entry frequency="3">scene</entry>
            <entry frequency="3">score</entry>
            <entry frequency="3">second</entry>
            <entry frequency="3">settlement</entry>
            <entry frequency="3">side</entry>
            <entry frequency="3">soldier</entry>
            <entry frequency="3">speaker</entry>
            <entry frequency="3">speech</entry>
            <entry frequency="3">stage</entry>
            <entry frequency="3">state</entry>
            <entry frequency="3">story</entry>
            <entry frequency="3">team</entry>
            <entry frequency="3">term</entry>
            <entry frequency="3">thing</entry>
            <entry frequency="3">tide</entry>
            <entry frequency="3">transaction</entry>
            <entry frequency="3">treaty</entry>
            <entry frequency="3">trial</entry>
            <entry frequency="3">trip</entry>
            <entry frequency="3">university</entry>
            <entry frequency="3">verdict</entry>
            <entry frequency="3">voice</entry>
            <entry frequency="3">water</entry>
            <entry frequency="3">wind</entry>
            <entry frequency="3">writing</entry>
            <entry frequency="3">youngest</entry>
            <entry frequency="3">youth</entry>
        </substitution>
    </mwu_substitution>
    <lexical_substitution>
        <substitution relation="NN" position="1">
            <entry id="0">
                <score>1.0009990009999001</score>
                <substitute>plant</substitute>
                <arguments>re-processing reprocessing kozloduy
                    hop sellafield rabta fertilizer sorghum
                    fgd resurrection plastics pesticide
                    chemicals petro-chemical poison treatment
                    chemical coca pilot power weapons crop
                    petroleum fuel gas steel assembly bulgaria
                    car complete</arguments>
            </entry>
            <entry id="1">
                <score>0.30052995707644664</score>
                <substitute>plants</substitute>
                <arguments>sorghum crop treatment coca
                    power fuel weapons chemical car</arguments>
            </entry>
            <entry id="2">
                <score>0.23305620873045516</score>
                <substitute>industry</substitute>
                <arguments>petro-chemical steel car chemical
                    power gas weapons</arguments>
            </entry>
            <entry id="3">
                <score>0.16719424968272895</score>
                <substitute>supplies</substitute>
                <arguments>gas fuel petroleum power weapons</arguments>
            </entry>
            <entry id="4">
                <score>0.13304422950069655</score>
                <substitute>company</substitute>
                <arguments>petroleum car chemical steel</arguments>
            </entry>
            <entry id="5">
                <score>0.10016112919338727</score>
                <substitute>factory</substitute>
                <arguments>fuel chemical car</arguments>
            </entry>
            <entry id="6">
                <score>0.10020619797659647</score>
                <substitute>failure</substitute>
                <arguments>crop complete power</arguments>
            </entry>
            <entry id="7">
                <score>0.10008730609398017</score>
                <substitute>production</substitute>
                <arguments>coca weapons car</arguments>
            </entry>
            <entry id="8">
                <score>0.10010419687839042</score>
                <substitute>project</substitute>
                <arguments>petro-chemical pilot power</arguments>
            </entry>
            <entry id="9">
                <score>0.10015644287179204</score>
                <substitute>supply</substitute>
                <arguments>gas bulgaria power</arguments>
            </entry>
            <entry id="10">
                <score>0.10001802901635107</score>
                <substitute>workers</substitute>
                <arguments>steel gas car</arguments>
            </entry>
        </substitution>
        <substitution relation="VO" position="0">
            <entry id="0">
                <score>1.0009990009999001</score>
                <substitute>plant</substitute>
                <arguments>tree vegetables trees seeds
                    culture instability bombs crops bomb
                    material growth</arguments>
            </entry>
            <entry id="1">
                <score>0.3646515821840497</score>
                <substitute>planted</substitute>
                <arguments>bomb bombs seeds trees</arguments>
            </entry>
        </substitution>
        <substitution relation="VO" position="1">
            <entry id="0">
                <score>1.0009990009999001</score>
                <substitute>plant</substitute>
                <arguments>close becomes build designed
                    visited declared down used keep</arguments>
            </entry>
            <entry id="1">
                <score>0.33302518737301345</score>
                <substitute>aircraft</substitute>
```

```
    <arguments>build down used</arguments>
</entry>
<entry id="2">
    <score>0.33312257727842143</score>
    <substitute>bank</substitute>
    <arguments>close build used</arguments>
</entry>
</substitution>
</lexical_substitution>
```

CHAPTER D

SOFTWARE SYSTEM SUMMARY

In this appendix we will simply list the components of the software developed during the course of this project. A full documentation would be beyond the scope of this thesis, but it is expected that the software will eventually become available for public use.

Some components (e.g. the parts-of-speech tagger) had been developed separately before this project and are not directly part of this system but rather independent modules.

D.1 Package corpus

This package contains the module for accessing indexed corpus data. The central class is `Corpus`, which has two sub-classes, `SingleCorpus` and `Corpora`. Through this abstraction a single corpus can be treated just the same as a collection of corpora. The remaining classes are used for indexing and the organisation of the lexical relations used for the usage patterns (see section 5.3).

Corpus	Base class for corpus access functionality
Corpora	Sub-class to deal with collections of corpora
SingleCorpus	Sub-class to deal with a single corpus file
WordListIndexer	For indexing tokens in the corpus file
ReIndex	For re-running the index
PositionLister	Interface for a class that receives token positions
IndexReader	For reading the index file
LexicalRelations	Interface for processing usage pattern triples
LexRel	Processing usage pattern triples
LexRelAggregator	Handling relations for multiple corpora
Relation	A single relation

D.2 Package io

The index data is stored in a compressed format described in Witten *et al.* (1994); the classes in this package are used to read and write data files in this format.

BitInputStream	Read a stream of bits
BitOutputStream	Write a stream of bits
BitStreamTest	Unit test
GammaWriter	Write integer values in gamma encoding
BufferedRandomAccessFile	Buffered version of java.io.RandomAccessFile

D.3 Package util

There are a number of general utility classes that were developed for the system; since these are not tied to the system directly they have been kept separate. There is also a sub-package which implements a basic sparse matrix.

SparseMatrix	Basic sparse matrix implementation
SparseMatrixTest	Unit test
LabelNode	Node representing a row/column starting point
MatrixNode	Data node in the sparse matrix

Args	Processing command-line arguments
Constants	Set-up information such as data paths
DoubleRef	Mutable double value
IntRef	Mutable int value
FreqFilter	A filter to cap a list at a certain percentage
FreqMap	A map to keep track of frequency information
FreqEntry	A single FreqMap entry
FreqMapTest	Unit test
Lemmatiser	Basic English lemmatiser (Harris, 1985)
LemmatiserTest	Unit test
LexicalDensity	A class to compute lexical density

Maths	Various mathematical functions (log2, mi, etc)
Matrix	A matrix interface
Multiplexer	Distribute its input to various files according to a keyword
MultiStore	A map to store multiple items with a single key
MultiStoreTest	Unit test
PositionLister	For indexing tokens in a text file
PreTokeniser	Preprocessing stage for tokenisation
ProgressBar	A progress bar for console applications
Stack	A strongly-typed Stack implementation
StackTest	Unit test
StringTokeniser	A speed-optimised variant of the standard StringTokenizer
StringTokeniserTest	Unit test
Table	A table data structure
TableTest	Unit test
Tense	A tense/aspect/voice recogniser
TenseTest	Unit test
Util	A collection of utility methods
UtilTest	Unit test
WordPositionReceiver	An interface for indexing classes
WordTree	A Trie data structure with word as nodes
WordTreeTest	Unit test

D.4 Package grammar

This package contains the implementation of a parser based on a transition network. This parser is also used to recognise grammar patterns (see section 5.4).

Network	A basic interface for a network
State	A node in the network
Arc	An abstract network arc
SynArc	A syntactic class arc
TokenArc	A word token arc
Chart	A chart implementing `Network`
Pattern	A grammar pattern
PatternTest	Unit test
Processor	Basic functions for parsing
Record	A class to store parsing history
TransitionNetwork	A basic RTN parser (Winograd, 1983)
Parser	A pattern grammar parser
ParserTest	Unit test

D.5 Package methods

Unlike the `util` package, the classes contained in this package are more closely linked into the system, and thus are of less general use. They mostly are relevant for the implemented linguistic procedures.

Chains	Implementation of the ‘chains’ procedure
ColligationChains	Implementation of ‘colligation’
Collocate	A single collocate
Collocations	A set of collocates
CompressRelations	A class to reduce the space needed by lexical relations
Concordances	A representation for concordance lines
Distribution	A class for calculating spread and coverage of words (4.1.2)
Frames	Implementation of ‘frames’ procedure
Span	A set of words around the node word
SpanTest	Unit test
TaggedSpan	`Span` with added POS-tags
Window	Base class for window around node word
WindowTest	Unit test
HanningWindow	A Hann(ing) window for collocation processing
RectWindow	A rectangular window for collocation processing
TriangWindow	A triangular window for collocation processing
SignificanceFunction	Base class for significance functions
MiScore	mi significance score
ObservedExpected	observed/expected score
TObs	modified observed/expected score
TScore	t-score significance score
RawFrequency	raw frequency significance score

InflAnalyser	For analysis of inflectional variation
LexicalGravity	Calculates lexical gravity
MeaningUnit	Represents a Unit of Meaning
MethodTest	Unit test
UnitsOfMeaning	Units of Meaning recognition procedure
WordRelator	Procedure for identifying lexical relations in a corpus

D.6 Package parser

This parser is a different implementation as the one from the `grammar` package, and is used for the processing of colligation (see section 5.2). It was originally developed as a PSG parser. The package also contains a dependency grammar parser, which is not used in the current version of the system.

ChartNGrams	A class to retrieve n-grams ('chains') from a chart
ChartParser	A chart parser
ChartTest	Unit test
Chart	A chart
Edge	A chart edge
Node	A chart node
DGParser	A dependency parser
DGrammar	A set of dependency rules
Rule	A dependency rule
RuleTest	Unit test

D.7 Package process

In this package are all the classes that perform the actual analysis. Most of them are simply wrappers around classes in the `method` package, with additional house-keeping functions to store the results in the correct XML format. The central class is `DoIt`, which coordinates the sequence in which the procedures are called. `Processor` is an abstract class which all implementations of procedures inherit to make sure the interface is uniform.

DoIt	Coordinating class
Basic	Basic lexical statistics
ChainProc	Chains
Colligations	Colligations
Colls	Collocations
FrameProc	Frames
GrammarPatterns	Grammar Patterns
Infl	Inflectional distribution
LemmaColls	Lemma Collocates
LexGrav	Lexical Gravity
LexIntSub	Usage Pattern substitution (6.3)
MWUSub	Multi-word unit substitution (6.4)
Processor	Abstract base class for analytic procedures
ProcUtil	Auxiliary methods for processing
Relation	A class to represent lexical relations
SynArgs	Processing usage patterns
UoM	Identify Units of Meaning

CHAPTER E

EVALUATION DATA FOR USAGE PATTERNS

Key

@M	missed relation
@E	erroneous relation
@C	correct relation

AS with Ian Smith, the verbal opposition of Mrs Thatcher to the prevailing 'winds of change' in their respective countries has conferred heroic status in the annals of Right-wingism everywhere .

@M	PN	to change
@C	NN	ian smith
@C	AN	verbal opposition
@C	NN	mrs thatcher
@C	AN	respective countries
@E	NN	their countries
@C	AN	heroic status
@C	VO	conferred status
@C	PN	with smith
@E	PN	to the
@C	PN	in countries
@C	PN	in annals
@C	VP	conferred in

They are both worshipped as political martyrs who did their best against impossible odds .

@C	AN	political martyrs
@C	AN	impossible odds
@E	SV	they are
@M	SV	they worshipped
@M	VO	did best
@E	SV	both worshipped
@C	PN	as martyrs
@C	VP	worshipped as
@M	VP	did against
@C	PN	against odds

Anyone holding a contrary view is dismissed as a prejudiced crank or, more gently, as one overdoing the conspiracy theory of history through misplaced zeal .

@M	SV	anyone holding
@M	SV	one overdoing
@M	VO	holding view
@M	AN	contrary view
@C	AN	prejudiced crank
@C	NN	conspiracy theory
@C	AN	misplaced zeal
@C	VO	overdoing theory|history
@C	PN	as crank
@C	VP	dismissed as
@C	PN	through zeal
@C	VP	overdoing through

The present writer figures in both categories, but remains of the same opinion still !

@E	NN	present figures
@E	NN	writer figures
@E	NN	both categories
@C	PN	in categories
@M	AN	present writer
@M	SV	writer figures
@M	AN	both categories
@M	AN	same opinion

"It is all of a piece," as the cook said in some forgotten thriller .

@C	AN	forgotten thriller
@C	SV	it is
@C	SV	cook said
@C	VO	is all|piece
@C	PN	in thriller
@C	VP	said in

The mysterious Power that purports to establish universal hegemony in defiance of the Creator's Word and will for fallen humanity's salvation and celestial destiny has pursued that aim by much the same method since our first parents' expulsion from Eden .

@C	AN	mysterious power
@C	AN	universal hegemony
@C	NN	creator word
@E	NN	's word
@E	AN	fallen salvation
@E	NN	humanity salvation
@E	NN	's salvation
@E	NN	our expulsion
@E	NN	parents expulsion
@E	NN	' expulsion
@C	SV	that purports
@E	SV	destiny pursued
@E	SV	that aim
@C	VO	establish hegemony
@C	PN	in defiance|word
@C	VP	establish in
@C	PN	for salvation
@C	PN	by much
@E	VP	aim by
@C	PN	from eden
@C	Vinf	purports establish

Always the ultimate attainment has been periodically frustrated, not by contemporaneous society but by the wrath of God .

@C	AN	ultimate attainment
@C	VO	frustrated attainment
@C	PN	by wrath|god
@M	AN	contemporareous society

So gradual and devious have been the recurring campaigns of successive pawns of the Devil in pursuit of his aspiration to world domination that nearly always "the little victims play," ignoring the evidence of approaching doom and disregarding any divinely inspired prophets sent to warn them .

@M	AN	approaching doom
@M	AN	inspired prophets
@C	AN	recurring campaigns
@M	VO	sent prophets
@C	AN	successive pawns
@E	NN	his aspiration
@C	NN	world domination
@C	AN	little victims
@E	SV	victims play
@E	SV	any divinely
@E	SV	prophets sent
@C	VO	been campaigns|pawns|devil
@C	VO	warn them
@C	PN	in pursuit|aspiration
@C	VP	been in
@C	PN	to domination
@E	VP	been to
@C	Vinf	sent warn

Emerging Pattern In 1529 England was still a Catholic kingdom, despite Luther, Calvin or whoever else set up as the latest light of the world .

@C	NN	catholic kingdom
@C	AN	latest light
@C	SV	pattern was
@C	VO	was kingdom
@C	PN	in england
@C	PN	as light|world
@C	VP	set as

Then the pattern began once more to emerge .

@C	SV	pattern began
@C	Vinf	began emerge

The ambitions of a lustful, profligate king and a cold-hearted wanton opened Pandora’s box .

@C	AN	cold-hearted wanton
@C	SV	wanton opened
@E	VO	opened pandora
@M	VO	opened box
@M	AN	profligate king
@M	SV	ambitions opened

In pursuit of his desires, Henry VIII forgot that he, like Pilate, would have no power if it "was not given him from above . "

@M	SV	Henry forget
@M	VP	given from
@E	NN	his desires
@E	SV	no power
@C	VO	forgot that
@C	VO	given him
@C	PN	in pursuit|desires

With his abrogation of the power given to the Vicar of Christ on earth, he made the monarchy subservient, in the first instance, to the receivers of the wealth stolen from the Church, and subsequently to the bankers of those same receivers .

@M	VO	given power
@M	VO	stolen wealth
@E	NN	his abrogation
@C	SV	he made
@C	VO	made monarchy
@C	PN	with abrogation|power
@C	PN	to vicar|christ
@C	VP	given to
@C	PN	on earth
@C	VP	given on
@C	PN	in instance
@C	PN	to receivers|wealth
@C	PN	from church
@C	VP	stolen from
@C	PN	to bankers

He murdered, more or less legally, the only two far-sighted enough to visualise the end result of his abrogation :

@C	NN	end result
@E	NN	his abrogation
@C	SV	he murdered
@C	VO	visualise result|abrogation

John Fisher, Bishop of Rochester, and Thomas More, erstwhile Chancellor of England .

@C	NN	john fisher
@M	NN	thomas more
@C	AN	erstwhile chancellor

The unity of Christendom was shattered into myriad sects .

@E	NN	myriad sects
@M	AN	myriad sects
@C	VO	shattered unity|christendom
@C	PN	into sects
@C	VP	shattered into

Faith and charity were in eclipse and even hope died at last .

@C	SV	charity were
@M	SV	hope died
@M	PN	in eclipse

If all the bishops had seen what the dire result their 'patriotic' obedience (if that is what it was) would have, would they have taken that first step on the long shuffle to the new paganism ?

@M	AN	first step
@M	VP	taken on
@M	VP	taken to
@M	VO	taken step
@M	PN	on to
@M	PN	on shuffle
@C	AN	dire result
@C	AN	new paganism
@C	SV	bishops seen
@C	SV	that is
@C	SV	it was
@C	SV	they taken
@E	VO	taken that
@E	PN	on the
@E	VP	step on
@C	PN	to paganism
@E	VP	shuffle to

Alas, they saw no harm in accepting their king as head of the Church in England .

@E	NN	their king
@C	SV	they saw
@E	SV	no harm
@C	PN	in england
@M	VO	accepting king

When in due time the Church in England became the Church of England, it was too late .

@M	AN	due time
@C	SV	church became
@C	SV	it was
@C	VO	became church|england
@C	PN	in england
@E	VP	time in

The lay Catholics of England and the 'hedge' priests died for their faith in their hundreds: the landed gentry were either martyred or fined into exile or apostasy .

@C	NN	lay catholics
@C	NN	hedge priests
@E	NN	' priests
@E	NN	their faith
@E	NN	their hundreds
@C	AN	landed gentry
@C	SV	priests died
@C	SV	gentry were
@C	PN	for faith
@C	VP	died for
@C	PN	in hundreds
@C	VP	died in
@C	PN	into exile
@C	VP	fined into

The 'new rich' clung to their ill-gotten gains through restorations and regicide, evolving into the die-soft Tory 'opposition' of today .

@M	AN	new rich
@M	SV	rich clung
@M	PN	through restorations
@M	PN	into opposition
@C	AN	ill-gotten gains
@E	NN	their gains
@C	PN	to gains
@C	VP	clung to
@E	PN	into die-soft

Usurers returned and by the end of the seventeenth century had annexed the royal monopoly of the issue of credit .

@C	AN	royal monopoly
@C	SV	usurers returned
@C	VO	annexed monopoly|issue|credit
@C	PN	by end|century

The Money Power ruled again .

@C	NN	money power
@C	SV	power ruled

It took four hundred years to turn Christendom into the multi-religious and multi-racial stew that is Europe today .

@C	NN	hundred years
@C	AN	multi-racial stew
@C	SV	it took
@C	SV	that is
@C	VO	took years
@C	VO	turn christendom
@C	VO	is europe
@C	PN	into multi-religious
@C	VP	turn into
@C	Ninf	years turn

It took rather less time to destroy the intervening imperial substitutes, for the labourers in the colonial vineyards were quite as gullible as their predecessors .

@C	AN	intervening substitutes
@C	AN	imperial substitutes
@C	AN	colonial vineyards
@E	NN	their predecessors
@C	SV	it took
@E	SV	less time
@C	VO	destroy substitutes
@C	PN	for labourers
@C	PN	in vineyards
@C	PN	as predecessors
@C	VP	gullible as
@C	Vinf	time destroy

BIBLIOGRAPHY

Aarts, B. (2000). Corpus linguistics, Chomsky and fuzzy tree fragments. In Mair and Hundt (2000), pages 5–13.

Allén, S. (1981). The lemma-lexeme model of the Swedish lexical data base. In B. B. Rieger, editor, *Empirical Semantics*, volume II, pages 376–387. Brockmeyer, Bochum.

Allén, S. (1982). *Text Processing: Proceedings of Nobel Symposium 51*. Almqvist & Wiksell International, Stockholm.

Altmann, G. (1980). Prolegomena to Menzerath's law. In R. Grotjahn, editor, *Glottometrika 2*, pages 1–10. Brockmeyer, Bochum.

Archangeli, D. and Langendoen, D. T., editors (1997). *Optimality Theory: An Overview*. Blackwell, Oxford.

Argamon, S., Akiva, N., Amir, A., and Kapah, O. (2004). Efficient unsupervised recursive word segmentation using minimum description length. In *Proceedings of COLING 2004*, Geneva.

Atwell, E. S. (1988). Transforming a parsed corpus into a corpus parser. In M. Kytö, O. Ihalainen, and M. Rissanen, editors, *Corpus Linguistics, Hard and Soft: Proccedings of the Eight International Conference on English Language Research on Computerized Corpora*, pages 61–69, Amsterdam. Rodopi.

Axelrod, R. and Cohen, M. D. (2000). *Harnessing Complexity: Organizational Implications of a Scientific Frontier*. The Free Press, New York.

Baayen, H. R. (1991). A stochastic process for word frequency distributions. In D. E. Appelt, editor, *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 271–278.

Baayen, H. R. (2001). *Word Frequency Distributions*. Kluwer, Dordrecht.

Baker, M., Francis, G., and Tognini-Bonelli, E., editors (1993). *Text and Technology*. John Benjamins, Amsterdam.

Barnbrook, G. (1996). *Language and Computers*. Edinburgh University Press, Edinburgh.

Barnbrook, G. (2002). *Defining Language: A local grammar of definition sentences*. John Benjamins Publishers, Amsterdam.

Barron, A., Rissanen, J., and Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, **6**(44), 2743–2760.

Berry-Rogghe, G. L. (1973). The computation of collocations and their relevance in lexical studies. In A. Aitken, R. Bailey, and N. Hamilton-Smith, editors, *The Computer and Literary Studies*, pages 102–112. EUP, Edinburgh.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman.

Black, E., Abney, S., Flickenger, D., Gdaniec, C., Grishman, R., Harrison, P., Hindle, D., Ingria, R., Jelinek, F., Klavans, J., Liberman, M., Marcus, M., Roukos, S., Santorini, B., and Strzalkowski, T. (1991). A procedure for quantitatively comparing the syntactic coverage of English grammars. In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*. Morgan Kaufman.

Black, E., Garside, R., and Leech, G., editors (1993). *Statistically-Driven Computer Grammars of English: The IBM/Lancaster Approach*. Rodopi, Amsterdam.

Bloomfield, L. (1933). *Language*. Holt, Rinehart and Winston, New York.

Brent, M. R. (1993). From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, **19**(2), 243–262.

Brill, E. and Resnik, P. (1994). A rule-based approach to prepositional phrase attachment disambiguation. 15th International Conference on Computational Linguistics (COLING94).

Briscoe, E. and Carroll, J. (1995). Towards automatic extraction of argument structure from corpora. ACQUILEX II Working Paper.

Chierchia, G. and McConnell-Ginet, S. (1990). *Meaning and Grammar: An Introduction to Semantics*. MIT Press, Cambridge, MA.

Chomsky, N. A. (1957). *Syntactic Structures*. Mouton, The Hague.

Chomsky, N. A. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Church, K. W. and Hanks, P. (1989). Word association norms, mutual information and lexicography. In *Proceedings of ACL 27*, pages 76–83.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**(1), 22–29.

Church, K. W., Gale, W., Hanks, P., and Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik, editor, *Lexical Acquisition: Using On-line Resources to Build a Lexicon*, pages 115–164. Lawrence Erlbaum, Englewood Cliffs, NJ.

Cobuild (1995). Collins COBUILD English collocations on CD-ROM.

Collins (1991). *Collins English Dictionary*. HarperCollins, Glasgow.

Covington, M. A. (1984). *Syntactic Theory in the High Middle Ages*. Cambridge University Press.

Covington, M. A. (1990). A dependency parser for variable-word-order languages. Technical report, The University of Georgia, Athens, GA. report AI-1990-01.

Creutz, M. and Lagus, K. (2002). Unsupervised discovery of morphemes. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, pages 21–30, Philadelphia.

Crystal, D. (1992). *An Encyclopedic Dictionary of Language and Languages*. Blackwell, Oxford.

Danielsson, P. (2001). *The automatic identification of meaningful units in language*. Ph.D. thesis, Göteborg University.

de Saussure, F. (1916). *Cours de Linguistique Generale*. Payot, Paris.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, **41**(6), 391–407.

Dias, G. (2003). Multiword unit hybrid extraction. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 41–48.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

Edmonds, P. (2002). Senseval: The evaluation of word sense disambiguation systems. *ELRA Newsletter*, **7**(3).

Elman, J. L. (1995). Language as a dynamical system. In R. F. Port and T. van Gelder, editors, *Mind as Motion: Explorations in the Dynamics of Cognition*, pages 195–223. MIT Press, Cambridge, MA. Online at http://crl.ucsd.edu/ elman/ Papers/ dynamics/ dynamics.html.

Esser, J. (2000). Corpus linguistics and the linguistic sign. In Mair and Hundt (2000), pages 91–101.

Fielding, R. (2000). *Architectural Styles and the Design of Network-based Software Architectures*. Ph.D. thesis, University of California, Irvine.

Firth, J. R. (1951). Modes of meaning. In *Papers in Linguistics 1934–1951*. Oxford University Press, London.

Firth, J. R. (1952). Linguistic analysis as a study of meaning. In F. Palmer, editor, *Selected Papers of J.R. Firth 1952–59*, pages 12–26. Longmans, London and Harlow.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. *Studies in Linguistic Analysis*, pages 1–32.

Francis, G. (1991). Nominal group heads and clause structure. *Word*, **42**, 144–156.

Francis, G. (1993). A corpus-driven approach to grammar: Principles, methods and examples. In Baker *et al.* (1993), pages 137–156.

Francis, G., Hunston, S., and Manning, E. (1996). *Grammar Patterns 1: Verbs*. HarperCollins, London.

Frege, F. L. G. (1884). *Die Grundlagen der Arithmetik: eine logisch-mathematische Untersuchung über den Begriff der Zahl*. W. Koebner, Breslau.

Fries, C. C. (1952). *The Structure of English*. Longmans, London.

Garside, R., Leech, G., and Sampson, G., editors (1987). *The Computational Analysis of English: A Corpus-Based Approach*. Longman, London.

Gieseking, K. (1993). *Synergetische Aspekte von Struktur und Dynamik der englischen Lexik*. Master's thesis, University of Trier.

Girju, R., Badulescu, A., and Moldovan, D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of HLT-NAACL 2003*, pages 1–8.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, **27**(2), 153–198.

Graddol, D., Cheshire, J., and Swann, J. (1987). *Describing Language*. Open University Press, Milton Keynes.

Green, G. M. and Morgan, J. L. (1996). *Practical Guide to Syntactic Analysis*. Number 67 in Lecture Notes. CSLI Publications, Stanford.

Greenacre, M. J. (1993). *Correspondence Analysis in Practice*. Academic Press, London.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Press, Boston, MA.

Grefenstette, G. and Tapanainen, P. (1994). What is a word, what is a sentence? problems of tokenization. In *Proceedings of COMPLEX '94*, pages 79–87.

Gross, M. (1982). Simple sentences: Discussion of Fred W. Householder's paper "analysis, synthesis and improvisation". In Allén (1982), pages 297–315.

Gross, M. (1993). Local grammars and their representation by finite automata. In Hoey (1993), pages 26–38.

Gross, M. (1997). The construction of local grammars. In E. Roche and Y. Schabes, editors, *Finite State Language Processing*, pages 329–354. Bradford, Cambridge, MA.

Halliday, M. A. K. (1961). Categories of the theory of grammar. *Word*, **17**, 241–292.

Halliday, M. A. K. (1966). Lexis as a linguistic level. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, and R. H. Robbins, editors, *In Memory of J. R. Firth*, pages 148–162. Longman, London.

Halliday, M. A. K. (1993). Quantitative studies and probabilities in grammar. In Hoey (1993), pages 1–25.

Halliday, M. A. K. and James, Z. L. (1993). A quantitative study of polarity and primary tense in the English finite clause. In J. M. Sinclair, M. Hoey, and G. Fox, editors, *Techniques of Description: Spoken and Written Discourse*, pages 32–66. Routledge, London.

Harold, E. R. (2002). XOM. http://www.xom.nu/.

Harris, M. D. (1985). *Introduction to Natural Language Processing*. Reston Publishing Co., Reston, VA.

Harris, Z. (1946). From morpheme to utterance. *Language*, **22**, 161–183.

Harris, Z. (1955). From phonemes to morphemes. *Language*, **31**(2), 190–222.

Hastings, P. M. (1994). *Automatic Acquisition of Word Meaning from Context*. Ph.D. thesis, University of Michigan.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING'92*.

Helbig, G. (1983). *Geschichte der neueren Sprachwissenschaft*. Westdeutscher Verlag, Opladen.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, Pittsburgh. ACL.

Hindle, D. and Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, **19**(1), 103–120.

Hoey, M., editor (1993). *Data, Description, Discourse*. HarperCollins, London.

Hoey, M. (1998). 'Introducing applied linguistics': 25 years on. Plenary paper in the 31st BAAL Annual Meeting.

Hoey, M. (2003). What's in a word? *MED Magazine*. http://www.macmillandictionary.com/ MED-Magazine/ August2003/ 10-Feature-Whats-in-a-word.htm.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An eternal golden braid*. Penguin, London.

Householder, F. W. (1982). Analysis, synthesis and improvisation. In Allén (1982), pages 271–295.

Huber, W. (1981). Semantic confusions in aphasia. In B. B. Rieger, editor, *Empirical Semantics II*, pages 423–445. Brockmeyer, Bochum.

Hunston, S. (2001). Colligation, lexis, pattern, and text. In M. Scott and G. Thompson, editors, *Patterns of Text*, pages 13–33. John Benjamins, Amsterdam.

Hunston, S. and Francis, G. (2000). *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. Benjamins, Amsterdam.

Justeson, J. S. and Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, **1**, 9–27.

Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A. (1995). *Constraint Grammar. A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter, Berlin and New York.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons, New York.

Kilgarriff, A. (1998). Senseval: An exercise in evaluating word sense disambiguation systems. In *Proceedings of LREC'98*, pages 581–588.

Kilgarriff, A. and Tugwell, D. (2001). Word sketch: Extraction and display of significant collocations for lexicography. In *Proceedings of the workshop 'Collocation: Extraction, Analysis and Exploitation, 39th ACL and 10th EACL, July 2001*, pages 32–38, Toulouse.

Kita, K., Kato, Y., Omoto, T., and Yano, Y. (1994). Automatically extracting collocations from corpora for language learning. In T. McEnery and A. Wilson, editors, *Corpora in Language Education and Research: A Selection of Papers from Talc94*. Department of Linguistics, Lancaster University.

Kittredge, R. I. (1981). Cohesive text structure in sublanguages. In B. B. Rieger, editor, *Empirical Semantics II*, pages 446–466. Brockmeyer, Bochum.

Kjellmer, G. (1984). Some thoughts on collocational distinctiveness. In J. Aarts and W. Meijs, editors, *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research*, pages 163–171. Rodopi, Amsterdam.

Kjellmer, G. (1987). Aspects of English collocation. In W. Meijs, editor, *Corpus Linguistics and Beyond*, pages 133–140, Amsterdam. Rodopi.

Knowles, G. and Don, Z. M. (2004). The notion of a 'lemma': Headwords, roots and lexical sets. *IJCL*, **9**(1), 69–81.

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Brockmeyer, Bochum.

Kuhn, T. (1963). *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago.

Leech, G. (1974). *Semantics*. Penguin, London.

Leech, G. (1997). Grammatical tagging. In R. Garside, G. Leech, and A. McEnery, editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pages 19–33. Longman, London.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL*, pages 768–774, Montreal.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Baker *et al.* (1993), pages 157–176.

Mair, C. and Hundt, M., editors (2000). *Corpus Linguistics and Linguistic Theory: Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20)*. Rodopi, Amsterdam.

Manning, C. D. and Schütze, H. (2000). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, **19**(2), 313–330.

Mason, O. (1996). Corpus access software: The CUE system. *Text Technology*, **6**(4), 257–266.

Mason, O. (1997). The weight of words: an investigation of lexical gravity. In B. Lewandowska-Tomaszczyk and P. J. Melia, editors, *Practical Applications of Language Corpora (PALC'97)*, pages 361–375. Łodz University Press, Łodz.

Mason, O. (2000a). A developer's view of corpus linguistics: the CUE system. In Mair and Hundt (2000), pages 233–241.

Mason, O. (2000b). Parameters of collocation: The word in the centre of gravity. In J. M. Kirk, editor, *Corpora Galore: Analyses and Techniques in Describing English*, pages 267–280. Rodopi, Amsterdam and Atlanta, GA.

Mason, O. (2000c). *Programming for Corpus Linguistics: How to Do Text Analysis with Java*. EUP, Edinburgh.

Mason, O. (2004). Automatic processing of local grammar patterns. In *Proceedings of 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*, pages 166–171, Birmingham.

Mason, O. and Hunston, S. (2004). The automatic recognition of verb patterns: A feasibility study. *International Journal of Corpus Linguistics*, **9**(2), 253–270.

Merkel, M. and Andersson, M. (2000). Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. In *Proceedings of RIAO'2000*, volume 1, pages 737–746. http://www.ida.liu.se/ magme/ publications/ merkel-andersson-riao-2000.pdf.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, **63**, 81–87.

Miller, G. A., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, **3**(4), 235–244.

Monaghan, J. (1979). *The Neo-Firthian Tradition and its Contribution to General Linguistics*. Max Niemeyer Verlag, Tübingen.

Nevin, B. E. (1992). Zellig S. Harris: An appreciation. *California Linguistic Notes*, **23**(2), 60–64.

Niedermair, G. T. (1986). Divided and valency-oriented parsing in speech understanding. In *COLING'86*, pages 593–595.

Nørretranders, T. (1998). *The User Illusion: Cutting Consciousness down to Size*. Allen Lanen (The Penguin Press), Harmondsworth.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh.

Okasha, S. (2002). *Philosophy of Science*. OUP, Oxford.

Pantel, P. and Lin, D. (2000). An unsupervised approach to prepositional phrase attachment using contextually similar words. In *Proceedings of ACL 2000, Hong Kong*, pages 101–108.

Paprotté, W. (1992). Korpuslinguistik–Rückkehr zum Strukturalismus oder Erneuerung der Computerlinguistik? *LDV-Forum*, **9**(2), 3–14.

Pearson, J. (1998). *Terms in Context*. Number 1 in Studies in Corpus Linguistics. John Benjamins, Amsterdam and Philadelphia.

Pedersen, T. and Chen, W. (1995). Lexical acquisition via constraint solving. In *Working Notes of the AAAI Spring Symposium on Representation and Acquisition of Lexical Knowledge*.

Pilch, H. (1976). *Empirical Linguistics*. Francke, München.

Pratchett, T., Stewart, I., and Cohen, J. (1999). *The Science of Discworld*. Ebury Press.

Quasthoff, U. (1998). Deutscher Wortschatz im Internet. *LDV-Forum*, **15**(2), 4–23.

Quemada, B. (1987). Notes sur la lexicographie et la dictionnairique. *Cahiers de Lexicologie*, **51**(2), 229–242.

Renouf, A. (2000). The time dimension in modern English corpus linguistics. In B. Kettemann and G. Marko, editors, *Teaching and Learning by Doing Corpus Analysis: Proceedings of TALC'04*, pages 27–41, Amsterdam. Rodopi.

Renouf, A. and Sinclair, J. M. (1991). Collocational frameworks in English. In K. Aijmer and B. Altenberg, editors, *English Corpus Linguistics*, pages 128–144. Longman, London.

Rieger, B. (1979). Repräsentativität: von der Unangemessenheit eines Begriffs zur Kennzeichnung eines Problems linguistischer Korpusbildung. In H. Bergenholtz and B. Schaeder, editors, *Textcorpora. Materialien für eine empirische Textwissenschaft*, pages 52–70. Scriptor, Kronberg.

Rieger, B. (1989). *Unscharfe Semantik*. Peter Lang, Frankfurt.

Ruge, G. (1992). Experiments on linguistically based term associations. *Information Processing & Management*, **28**(3), 317–332.

Ruge, G. (1997). Automatic detection of thesaurus relations for information retrieval applications. In *Foundations of Computer Science: Potential – Theory – Cognition*, pages 499–506.

Sampson, G. (1980). *Schools of Linguistics: Competition and evolution*. Hutchinson, London.

Sampson, G. (1987). Evidence against the grammatical/ungrammatical distinction. In W. Meijs, editor, *Corpus Linguistics and Beyond*, pages 219–226. Rodopi, Amsterdam.

Sampson, G. (1995). *English for the Computer*. Clarendon Press, Oxford.

Sampson, G. (2001). *Empirical Linguistics*. Continuum, London, New York.

Schone, P. and Jurafsky, D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of Empirical Methods in Natural Language Processing*, Pittsburgh, PA. http://www.stanford.edu/ jurafsky/ emnlp_2001_mwu_iii.pdf.

Schütze, C. T. (1996). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. The University of Chicago Press, Chicago.

Schütze, H. (1993). Part-of-speech induction from scratch. In *Proceedings of ACL 31*, pages 251–258.

Schwarz, C., Ruge, G., and Warner, A. J. (1991). Effectiveness and efficiency in natural language processing for large amounts of text. *Journal of the American Society for Information Science*, **42**(6), 450–456.

Sells, P. (1985). *Lectures on Contemporary Syntactic Theories*. CSLI Lecture Notes. University of Chicago Press, Stanford, CA.

Silberztein, M. (1993). *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Masson, Paris.

Sinclair, J., Jones, S., and Daley, R. (2004). *English Collocation Studies: The OSTI Report*. Continuum, London. Originally published 1970; this edition edited by Ramesh Krishnamurthy.

Sinclair, J. M. (1966). Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, and R. H. Robbins, editors, *In Memory of J. R. Firth*, pages 410–430. Longman, London.

Sinclair, J. M., editor (1987). *Looking Up: The Cobuild Project in Lexical Computing*. Collins, London.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Sinclair, J. M. (1996a). Preliminary reccomendations on corpus typology. Technical report, EAGLES. EAG–TCWG–CTYP/P.

Sinclair, J. M. (1996b). The search for units of meaning. *Textus*, **9**, 75–106.

Sinclair, J. M., editor (2001). *English Dictionary for Advanced Learners*. HarperCollins, London.

Sinclair, J. M. and Hunston, S. (2000). A local grammar of evaluation. In S. Hunston and G. Thompson, editors, *Evaluation in Text: Authorial Stance and the Construction of Discourse*, pages 74–101. Oxford University Press, Oxford.

Sinclair, J. M. and Jones, S. (1974). English lexical collocations: a study in computational linguistics. *Cahiers de Lexicologie*, **24**(2), 15–61.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, **19**(1), 143–177.

Steiner, P. (2004). *Wortarten und Korpus: Automatische Wortartenklassifikation durch distributionelle und quantitative Verfahren*. Shaker Verlag, Aachen.

Stubbs, M. (1993). British traditions in text analysis–from Firth to Sinclair. In M. Baker, G. Francis, and E. Tognini-Bonelli, editors, *Text and Technology. In Honour of John Sinclair*, pages 1–33. Benjamins, Amsterdam.

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, **2**(1), 23–55.

Stubbs, M. (1996). *Text and Corpus Analysis*. Blackwell, Oxford.

Stubbs, M. (2001). *Words and Phrases: corpus studies of lexical semantics*. Blackwell, Oxford.

Stubbs, M. (2003). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*, **7**(2), 215–244.

Stubbs, M. and Barth, I. (2003). Using recurrent phrases as text-type discriminators. *Functions of Language*, **10**(1), 61–104.

Sudnow, D. (1978). *Ways of the Hand: The Organization of Improvised Conduct*. Harvard University Press.

Tesnière, L. (1959). *Eléments de Syntaxe Structurale*. Klincksieck, Paris.

Teubert, W. (1999). Corpus linguistics: a partisan view. *TELRI Newsletter*, **8**, 4–19.

Thurmair, G. (1984). Linguistic problems in multilingual morphological decomposition. In *Proceedings of the 10th International COLING Conference*, pages 174–177.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Studies in Corpus Linguistics. John Benjamins, Amsterdam.

Tretiakoff, A. (1973). Results obtained with a new method for the automatic analysis of sentence structures. In A. Zampolli and N. Calzolari, editors, *Computational and Mathematical Linguistics*, pages 215–233. Leo S. Olschki Editore, Florence.

Tufis, D. and Mason, O. (1998). Tagging Romanian texts: a case study for Qtag, a language independent probabilistic tagger. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, pages 589–596, Granada, Spain.

URL (2005). Online etymological dictionary. http://www.etymonline.com, last visisted 19/04/2005.

Willis, D. (1993). Grammar and lexis: Some pedagogical implications. In J. M. Sinclair, M. Hoey, and G. Fox, editors, *Techniques of Description: Spoken and Written Discourse*, pages 83–93. Routledge, London and New York.

Wilson, G., editor (1967). *A Linguistics Reader*. Harper & Row, New York.

Winograd, T. (1983). *Language as a cognitive process*. Addison-Wesley, Reading, MA.

Witten, I., Moffat, A., and Bell, T. C. (1994). *Managing Gigabytes: Compressing and Indexing Documents and Images*. Van Nostrand Reinhold, New York.

Wittgenstein, L. (1921). Logisch-philosophische Abhandlung. *Annalen der Naturphilosophie*, **14**.

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, **8**, 338–353.

Zampolli, A. (1994). Introduction. In B. T. S. Atkins and A. Zampolli, editors, *Computational Approaches to the Lexicon*, pages 3–15. Oxford University Press, Oxford.

Zipf, G. K. (1935). *The Psychobiology of Language*. Houghton Mifflin, Boston, MA.